Ай-ТИ Фреш

ТЕХНИЧЕСКИЙ РАЗБОР

Ceph: распределённое хранилище для корпоративной инфраструктуры

Наша методология: cephadm-развёртывание Ceph 20.x Tentacle — от подбора железа до эксплуатации

Июль 2026

itfresh.ru · ИТ-аутсорсинг для юридических лиц

О ЧЁМ ЭТОТ РАЗБОР

Суть проблемы

Файлы, базы 1С и образы ВМ у клиентов обычно живут на одном сервере с RAID: отказ контроллера или переполнение тома останавливает всю компанию, а расширение — это простой и миграция на новую СХД. Мы решаем это кластером Ceph: данные размазаны по узлам, хранилище переживает отказ диска и целого сервера без остановки сервисов и расширяется добавлением узла в рабочее время.

Почему это важно бизнесу

- Отказ одиночной СХД останавливает сразу всё: 1С, файлы, почту, ВМ — простой всей компании, а не одного сервиса
- Ceph — open source без лицензий за терабайт: платим только за железо, растём серверами стандартной комплектации
- Ёмкость расширяется без миграций и остановок: добавили узел — кластер сам перераспределил данные
- Замена диска и обновление узлов идут без окна простоя: сервисы работают, пока кластер восстанавливает избыточность
- Один кластер закрывает три задачи: диски ВМ (RBD), файловые шары (CephFS) и S3-хранилище под бэкапы (RGW)

ЧТО МЫ ЗАКЛАДЫВАЕМ

Ключевые параметры реализации

20.2.x

Ветка Ceph Tentacle — актуальный стабильный релиз; в прод берём point-релизы 20.2.2+, Reef (18...

docs.ceph.com, Ceph Releases (index), 2026

3 / 2

size/min_size реплицированных пулов: переживаем отказ целого узла без остановки записи и без р...

docs.ceph.com, Pools (v20.2)

0.85

nearfull_ratio — предупреждение при 85% заполнения OSD (меняется ceph osd set-nearfull-ratio);...

docs.ceph.com, OSD Config Reference (v20.2)

4 ГиБ

osd_memory_target на каждый OSD-демон; плюс закладываем ~1 ГБ RAM на 1 ТБ данных под пики reco...

docs.ceph.com, Hardware Recommendations (v20.2)

50 мс

mon_clock_drift_allowed (0.05 с) — порог рассинхрона часов, дальше MON_CLOCK_SKEW и риск квору...

docs.ceph.com, Monitor Config Reference (v20.2)

2×25GbE

LACP-бонд и раздельные public_network / cluster_network: recovery-трафик не душит клиентский I...

наш сетевой стандарт ITfresh, 2026

РЕАЛИЗАЦИЯ 1 ИЗ 3

Фундамент кластера: узлы, сеть и bootstrap через cephadm

Что настраиваем

5 однотипных узлов (MON+MGR+OSD), Debian 12, NVMe под BlueStore WAL/DB, изолированные сети public и cluster

Как мы это делаем

1. Готовим узлы: chrony, swap off, kernel.pid_max=4194303; под OSD — только enterprise-диски с PLP, контроллер в HBA/IT-mode, без RAID
2. cephadm bootstrap --mon-ip на первом узле; раздаём /etc/ceph/ceph.pub, добавляем узлы ceph orch host add; 5 MON и 2 MGR через ceph orch apply
3. OSD описываем drive-спецификацией service_type: osd (data_devices: NVMe, db_devices для гибрида) вместо --all-available-devices — состав дисков под контролем
4. Сети: public_network и cluster_network в конфиге, LACP-бонд 2×25GbE в отдельных VLAN; MTU 9000 проверяем ping -M do -s 8972 между всеми узлами

РЕЗУЛЬТАТ

Кластер поднимается в контейнерах за один рабочий день, оркестратор сам следит за демонами и перезапускает их. Добавление узла — одна команда ceph orch host add: диски подхватываются по drive-спецификации, данные перераспределяются автоматически.

КЛЮЧЕВОЙ НЮАНС

cephadm управляет только тем, что описано в service-спецификациях: держим их в git и применяем через ceph orch apply -i, а не правим демоны руками — иначе оркестратор молча вернёт старую конфигурацию.

РЕАЛИЗАЦИЯ 2 ИЗ 3

Пулы данных: репликация под ВМ, erasure coding под бэкапы

Что настраиваем

RBD-пул для Proxmox, CephFS для файловых шар, RGW (S3) для бэкапов — три вида доступа на одном кластере

Как мы это делаем

1. RBD: реплицированный пул size=3/min_size=2, pg_autoscale_mode on + target_size_ratio, rbd pool init; в Proxmox — storage типа RBD с отдельным keyring, не admin
2. Холодные данные: EC-профиль k=4 m=2 crush-failure-domain=host, allow_ec_overwrites true; метаданные EC-пулов всегда на replicated-пуле
3. CephFS: ceph fs volume create, MDS active + standby-replay (allow_standby_replay true) — фейловер метаданных за секунды, а не минуты
4. RGW: два инстанса за HAProxy, пользователи через radosgw-admin user create, ключи доступа — в парольном хранилище, бакеты по least privilege
5. Включаем балансировщик: ceph balancer mode upmap; ceph balancer on — PG распределяются равномерно без ручного reweight

РЕЗУЛЬТАТ

Один кластер закрывает блочный, файловый и объектный доступ: репликация даёт быстрые диски ВМ (33% полезной ёмкости), EC 4+2 — экономичные бэкапы (66% полезной ёмкости). Отказ любого узла не останавливает ни ВМ, ни шары, ни выгрузку бэкапов.

КЛЮЧЕВОЙ НЮАНС

size=2/min_size=1 в проде не используем никогда: одиночный сбой ведёт к inactive PG, двойной — к потере данных. EC медленнее на random write даже с allow_ec_overwrites — под горячие ВМ только репликация.

РЕАЛИЗАЦИЯ 3 ИЗ 3

Мониторинг, пороги ёмкости и регламент эксплуатации

Что настраиваем

MGR-модуль prometheus (:9283) → наш мониторинг с алертами в Telegram; дашборд MGR (:8443) только через VPN

Как мы это делаем

1. ceph mgr module enable prometheus; снимаем ceph_health_status, ceph_osd_up/ceph_osd_in, ceph_pg_active против ceph_pg_total, ceph_cluster_total_used_bytes/ceph_clus…
2. Пороги ёмкости: наш алерт планирования — 75%, дальше штатные nearfull 0.85, backfillfull 0.90, full 0.95; до full не доводим — запись встанет
3. QoS: планировщик mclock_scheduler, профиль high_client_ops в рабочие часы; scrub и deep-scrub — только в ночное окно через osd_scrub_begin_hour/osd_scrub_end_hour
4. Регламент: еженедельно ceph crash ls и ceph health detail, бэкап monmap (ceph mon getmap -o) и /etc/ceph с ключами вне кластера, раз в квартал — учебный вывод OSD

РЕЗУЛЬТАТ

Дежурный видит деградацию раньше пользователей: упавший OSD, degraded PG, приближение к nearfull. Типовой инцидент «умер диск» проходит без участия клиента — recovery автоматический, мы только меняем железо в плановом порядке.

КЛЮЧЕВОЙ НЮАНС

Смотрим заполнение самого полного OSD, а не среднее по кластеру: перекос в 10–15% — норма, и именно самый полный диск первым упрётся в full и остановит запись всего пула.

ЧТО МЫ ВЫУЧИЛИ НА ПРАКТИКЕ

Подводные камни

✕ size=2 / min_size=1 в продакшене

Экономия трети сырой ёмкости оборачивается inactive PG при одиночном сбое и потерей данных при двойном. Только size=3/min_size=2 или EC с m≥2.

✕ Consumer-SSD без PLP под OSD

BlueStore пишет WAL синхронно: без Power Loss Protection латентность fsync растёт в десятки раз, ресурс сгорает за месяцы. Только DC-серии NVMe/SSD.

✕ OSD поверх RAID-контроллера

Ceph сам обеспечивает избыточность, а RAID прячет SMART и латентность конкретного диска. Контроллер переводим в HBA/IT-mode, один OSD на один диск.

✕ Заполнение кластера выше 85%

После nearfull деградирует производительность, при full (0.95) запись останавливается полностью. Мы расширяемся с 75% и следим за самым полным OSD.

✕ Рассинхрон времени на MON

Paxos-кворум чувствителен к skew: уже 50 мс дают MON_CLOCK_SKEW и риск развала кворума. chrony на всех узлах от одного источника времени.

✕ Recovery душит клиентский I/O

При отказе диска rebalance забивает общую сеть. Отдельная cluster_network и mClock-профиль high_client_ops держат латентность ВМ в норме.

✕ Единственный MDS для CephFS

Отказ единственного MDS останавливает файловую систему на минуты. Ставим standby-replay — горячий резерв повторяет журнал активного MDS.

✕ PG-count, посчитанный наугад

Мало PG — перекос данных по OSD, много — лишний расход RAM/CPU. Включаем pg_autoscaler с target_size_ratio и флагом bulk для больших пулов.

УРОВНИ РЕАЛИЗАЦИИ

Как правильно

МИНИМУМ

- 3 узла, size=3/min_size=2, failure-domain=host, 10GbE под кластерную сеть
- Enterprise-SSD с PLP под OSD и WAL/DB, контроллер в HBA-режиме
- chrony на всех узлах; алерты на HEALTH_WARN и заполнение 75%

НОРМАЛЬНО

- 5 узлов, LACP 2×25GbE, public и cluster сети разнесены по VLAN
- EC 4+2 под бэкапы, репликация под BM; balancer в режиме upmap
- MGR prometheus → внешний мониторинг, алерты в Telegram дежурному
- Service-спецификации cephadm в git, бэкап monmap и ключей вне кластера

ХОРОШО

- MDS standby-replay, RGW ×2 за HAProxy, дашборд MGR только через VPN
- rbd-mirror или вторая RGW-зона на резервной площадке для DR
- Квартальные учения: вывод узла из кластера с замером времени recovery
- mClock-профили по расписанию, deep-scrub только в ночные окна

ПРОВЕРЬТЕ У СЕБЯ

Чек-лист самопроверки

- [ ] Переживает ли кластер отказ целого узла без остановки записи (size=3, min_size=2, failure-domain=host)?
- [ ] Выделена ли отдельная cluster-сеть, чтобы recovery не тормозил рабочие ВМ и базы?
- [ ] Стоят ли под OSD только диски с Power Loss Protection, а контроллер — в HBA-режиме?
- [ ] Настроен ли алерт на 75% заполнения и виден ли самый заполненный OSD, а не средний процент?
- [ ] Синхронизировано ли время всех узлов через chrony и есть ли алерт на clock skew мониторов?
- [ ] Лежат ли service-спецификации cephadm, monmap и ключи /etc/ceph в резервной копии вне кластера?
- [ ] Есть ли standby-MDS для CephFS и второй RGW-инстанс за балансировщиком?
- [ ] Проводился ли за последний год учебный вывод диска или узла с замером времени восстановления?
- [ ] Работает ли кластер на поддерживаемой ветке (Squid 19.2.x или Tentacle 20.2.x), а не на EOL-релизе?

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 ёмкости, сети и состава узлов под Ceph — оценка за 1 день
- Развёртывание под ключ: cephadm, пулы RBD/CephFS/RGW, интеграция с Proxmox и S3-бэкапами
- Перенос данных с одиночных RAID-серверов и NAS на кластер без остановки сервисов
- Мониторинг 24/7: метрики MGR-модуля prometheus, алерты в Telegram, регламент scrub и recovery
- Сопровождение: обновления между релизами, расширение ёмкости узлами, учения по отказам

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 — v20.2)

02 Hardware Recommendations (docs.ceph.com — v20.2)

03 Pools, Placement Groups and CRUSH (docs.ceph.com — v20.2)

04 Erasure Code Profiles (docs.ceph.com — v20.2)

05 mClock Config Reference (OSD QoS) (docs.ceph.com — v20.2)

06 Prometheus Module (ceph-mgr) (docs.ceph.com — v20.2)

07 Ceph Releases (index): Tentacle / Squid (docs.ceph.com — 2026)

08 Наш шаблон service-спецификаций cephadm (itfresh.ru — 2026)

Основано на официальной документации продуктов и нашей практике внедрения.