Ай-ТИ Фреш

ТЕХНИЧЕСКИЙ РАЗБОР

Ceph в среднем бизнесе: когда кластер оправдан, а когда нет

Критерии выбора, экономика и границы применимости: Ceph против ZFS, DRBD и фирменных СХД

Июль 2026

itfresh.ru · ИТ-аутсорсинг для юридических лиц

О ЧЁМ ЭТОТ РАЗБОР

Суть проблемы

Компания растёт, объём данных растёт на 30%+ в год, а хранилище — либо стареющая фирменная СХД с дорогой поддержкой и долгими запчастями, либо разрозненные локальные диски гипервизоров. Отказ контроллера или конец поддержки означает простой всей виртуализации. Мы разбираем, когда ответом становится Ceph-кластер, а когда он — дорогая ошибка, и даём числовые критерии решения.

Почему это важно бизнесу

- Простой СХД останавливает сразу все VM: 1С, почту, файлы — час простоя офиса на 50 мест стоит дороже месяца поддержки хранилища
- Продление поддержки фирменной СХД — сотни тысяч в год; Ceph на б/у x86 даёт 2–3× экономии, но только при правильном масштабе
- Неверный выбор дорог в обе стороны: Ceph «на вырост» съест админ-часы, а одиночная СХД упрётся в потолок в разгар роста
- Кластер переживает отказ целого сервера без остановки бизнеса — страховка, которую видно только в момент аварии

ЧТО МЫ ЗАКЛАДЫВАЕМ

Ключевые параметры реализации

20.2

Ветка Tentacle — ставим её в новые кластеры; Squid 19.2 доводим до EOL (октябрь 2026) и апгрей…

Ceph Tentacle, docs.ceph.com

4 узла

Минимум узлов в наших проектах: потеря одного не рушит кворум MON и оставляет место ребалансу

наш стандарт

10/25G

Сеть: 10G — нижний порог по докам, мы закладываем 25G и делим public/cluster по VLAN

по докам Ceph (hardware)

4 ГиБ

osd_memory_target на каждый OSD-демон — RAM узла считаем от числа дисков, не наоборот

BlueStore, docs.ceph.com

85%

Порог nearfull (mon_osd_nearfull_ratio): выше него нет места на ребаланс при отказе узла

наш стандарт по докам Ceph

×3

Реплика size=3/min_size=2 для VM-дисков: usable = raw/3 — бюджет считаем честно от этого

наш стандарт

Аудит применимости: профиль нагрузки и матрица критериев

Что настраиваем

Вся ИТ-инфраструктура заказчика: гипервизоры, файловые шары, бэкапы — до закупки железа

Как мы это делаем

1. Снимаем профиль: iostat/fio на текущих хостах — IOPS, латентность, доля случайной записи; объём и рост считаем по каталогам бэкапов за 12 месяцев
2. Прогоняем матрицу: ≥4 сервера под хранилище, ≥20 ТБ данных, рост от 30%/год, нужны ≥2 типа доступа (RBD для VM + CephFS или S3-RGW)
3. Проверяем сеть: есть ли 10G+ и возможность выделить cluster-VLAN; без этого Ceph не согласуем
4. Оцениваем команду: Linux-админ и 4–8 ч/мес на эксплуатацию — свои или наш аутсорс
5. Выносим вердикт: Ceph / ZFS-репликация / DRBD / одиночная СХД — с расчётом TCO на 5 лет

РЕЗУЛЬТАТ

Заказчик получает обоснованное «да/нет» до трат на железо: либо смету кластера с расчётом ёмкости, либо более дешёвую альтернативу. Так мы отсекаем проекты, где Ceph стал бы чемоданом без ручки.

КЛЮЧЕВОЙ НЮАНС

Главный фильтр — не объём, а сочетание роста и типов доступа: если нужен только один пул VM-дисков без роста, ZFS-репликация закрывает задачу на порядок дешевле

РЕАЛИЗАЦИЯ 2 ИЗ 3

Сравнение с альтернативами: ZFS, DRBD, фирменная СХД

Что настраиваем

Пороговые точки, на которых мы переключаем проект с одной технологии на другую

Как мы это делаем

1. 1–2 гипервизора, до 15 VM: ZFS mirror/raidz2 + асинхронная репликация zfs send/recv (sanoid/syncoid), RPO 5–15 минут
2. 2 узла и нужна синхронная реплика: DRBD 9 с третьим diskless-узлом-арбитром под кворум; проще, чем Ceph, но один том и без scale-out
3. Фирменная СХД: считаем не цену коробки, а поддержку и запчасти на 5 лет; vendor lock-in и предел расширения фиксируем в смете
4. Ceph берём с 4+ узлов: единая плоскость под RBD+CephFS+RGW, расширение добавлением узла без остановки, отказ узла переживается штатно

РЕЗУЛЬТАТ

Каждая технология получает нишу с числовыми границами — решение перестаёт быть вкусовщиной. Клиент видит, в какой точке роста ему станет выгодно мигрировать на Ceph, и закладывает это в бюджет заранее.

КЛЮЧЕВОЙ НЮАНС

DRBD и ZFS не «хуже» — они дешевле в своей нише; Ceph выигрывает, только когда нужны одновременно scale-out, несколько типов доступа и переживание отказа целого узла

РЕАЛИЗАЦИЯ 3 ИЗ 3

Расчёт ёмкости и отказоустойчивости до закупки

Что настраиваем

Sizing узлов и пулов: реплика, erasure coding, пороги заполнения — на этапе сметы

Как мы это делаем

1. Usable для VM-пулов: raw × 0.85 / 3 (size=3, min_size=2, потолок — nearfull 0.85); только это число показываем заказчику как ёмкость
2. EC-профиль под бэкапы: k=4 m=2 при crush-failure-domain=host требует 6 хостов; на 5 узлах берём k=3 m=2 — экономия места ~45% против реплики ×3
3. RAM узла: osd_memory_target 4 ГиБ × число OSD + 16 ГБ базы; HDD-OSD — только с NVMe под WAL+DB (DB 1–4% объёма OSD: RBD 1–2%, RGW от 4%)
4. Закладываем headroom: кластер должен вместить данные при выпадении одного узла и остаться ниже backfillfull 0.90

РЕЗУЛЬТАТ

Смета железа сходится с реальной ёмкостью: нет сюрприза «купили 600 ТБ raw — получили 170 usable». Кластер переживает отказ узла без остановки записи и без ночных авралов с ребалансом.

КЛЮЧЕВОЙ НЮАНС

Ошибка №1 в чужих сметах, которые мы переделываем: ёмкость посчитана от raw без учёта репликации и nearfull-порога — реальная usable-ёмкость в 3,5 раза меньше raw

ЧТО МЫ ВЫУЧИЛИ НА ПРАКТИКЕ

Подводные камни

- **Три узла «для экономии»**
  Кворум MON держится, но после отказа узла size=3 негде восстановить третью копию. Берём минимум 4 узла — деградация лечится ребалансом

- **EC 4+2 на пяти хостах**
  При crush-failure-domain=host профилю нужно k+m=6 доменов — PG останутся undersized. На 5 узлах используем k=3 m=2 или реплику

- **Сеть 1G «на первое время»**
  Каждая запись едет по сети трижды (size=3): на 1G латентность убивает VM-диски. Минимум 10G, public и cluster — раздельными VLAN

- **HDD-OSD без NVMe под WAL+DB**
  RocksDB и WAL на HDD роняют случайную запись в разы. Выносим на NVMe: DB 1–4% объёма OSD (RBD 1–2%, RGW от 4%), один NVMe на 4–6 HDD

- **Заполнение выше 85%**
  После nearfull (0.85) нет места на ребаланс, на full (0.95) кластер останавливает запись. Закупку следующего узла планируем при 70–75%

- **min_size=1 ради аптайма**
  Запись при одной живой реплике — потеря данных при втором отказе. Держим size=3/min_size=2 и не снижаем даже в инцидентах

- **Кластер без владельца**
  Ceph деградирует тихо: scrub-ошибки, копящиеся PG-варнинги. Закладываем 4–8 ч/мес админа или аутсорс-контракт — иначе не берёмся

- **Смета от raw-ёмкости**
  Заказчику обещают raw, а usable = raw × 0.85 / 3. В КП показываем только usable по каждому пулу с учётом его профиля

УРОВНИ РЕАЛИЗАЦИИ

Как правильно

МИНИМУМ

- 4 узла, size=3/min_size=2, MON×3, MGR×2; развёртывание через cephadm
- Сеть 10G, public и cluster в разных VLAN
- NVMe под WAL+DB для каждого HDD-OSD; RAM: 4 ГиБ на OSD + 16 ГБ
- Версия не ниже Squid 19.2 с планом апгрейда до Tentacle 20.2

НОРМАЛЬНО

- 5 узлов: отказ одного переживаем с полным восстановлением избыточности
- 25G LACP на два коммутатора (MLAG), DAC-линки
- Prometheus + ceph-exporter + Grafana; алерты: OSD down >5 мин, PG inactive, nearfull
- EC-пул k=3 m=2 под бэкапы, replicated — под VM-диски

ХОРОШО

- 6+ узлов — открывается EC 4+2 с failure-domain=host для архивов
- Тестовый стенд на VM для обкатки минорных и мажорных апгрейдов до прода
- Второй контур бэкапов вне кластера: RGW → off-site S3 или PBS на отдельном железе
- Регламент ёмкости: закупка следующего узла при 70% заполнения

ПРОВЕРЬТЕ У СЕБЯ

Чек-лист самопроверки

- [ ] Есть ли минимум 4 сервера под хранилище и 10G-сеть с отдельным cluster-VLAN?
- [ ] Посчитана ли usable-ёмкость от raw с учётом size=3 и nearfull 0.85, а не «объём дисков»?
- [ ] Проверено ли, что EC-профиль (k+m) не превышает число хостов при failure-domain=host?
- [ ] Хватает ли RAM: 4 ГиБ на каждый OSD-демон плюс 16 ГБ на систему узла?
- [ ] Есть ли у HDD-OSD выделенный NVMe под WAL+DB (DB 1–4% объёма OSD)?
- [ ] Назначен ли ответственный за кластер и заложены ли 4–8 ч/мес на эксплуатацию?
- [ ] Настроены ли алерты: OSD down дольше 5 минут, PG inactive, приближение к nearfull?
- [ ] Есть ли план апгрейда до актуальной ветки (Tentacle 20.2) и стенд для его проверки?
- [ ] Хранится ли хотя бы один контур бэкапов вне Ceph-кластера?
- [ ] Сравнён ли TCO Ceph с ZFS-репликацией и DRBD для вашего масштаба до закупки железа?

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 применимости: профиль нагрузки, матрица критериев, вердикт Ceph/ZFS/DRBD/СХД с TCO на 5 лет
- Sizing и смета: расчёт usable-ёмкости, узлов, сети и RAM под ваш рост; подбор б/у серверного железа
- Развёртывание под ключ: cephadm, пулы под VM/файлы/S3, интеграция с Proxmox, мониторинг с алертами
- Эксплуатация по SLA: апгрейды версий, замена дисков, контроль ёмкости и здоровья кластера

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 Hardware Recommendations (docs.ceph.com — Tentacle)

02 Ceph Releases (index): Tentacle 20.2, Squid 19.2 (docs.ceph.com — 2026)

03 BlueStore Configuration Reference (osd_memory_target, WAL/DB sizing) (docs.ceph.com — 20.2)

04 Erasure Code Profiles (crush-failure-domain) (docs.ceph.com — 20.2)

05 Monitor Config Reference — Storage Capacity (mon_osd_nearfull_ratio, full… (docs.ceph.com — 20.2)

06 Наш шаблон аудита СХД и матрица выбора хранилища (itfresh.ru — 2026)

Основано на официальной документации продуктов и нашей практике внедрения.