Ай-ТИ Фреш

ТЕХНИЧЕСКИЙ РАЗБОР

Proxmox VE: кластер виртуализации за один день

Наша методология: 3 ноды PVE 9.2, Ceph/ZFS, HA и бэкап в PBS — от голого железа до боевых VM

Июль 2026

itfresh.ru · ИТ-аутсорсинг для юридических лиц

О ЧЁМ ЭТОТ РАЗБОР

Суть проблемы

У заказчика на 30–50 рабочих мест сервисы (1С, SQL, файлы, CRM) живут на 2-3 одиночных хостах: падение любого = простой всего офиса на часы, а продление лицензий импортного гипервизора стало непредсказуемым по цене и срокам. Мы за один рабочий день собираем кластер Proxmox VE с общим хранилищем и HA: отказ ноды переживается автоматическим рестартом VM за минуты.

Почему это важно бизнесу

- Отказ одиночного хоста с 1С и файлами останавливает работу всех сотрудников до ремонта железа
- Лицензии зарубежных гипервизоров не продлеваются штатно — риск остаться без обновлений и поддержки
- Proxmox VE — AGPLv3: ноль лицензионных платежей за гипервизор, HA, живую миграцию и бэкап
- Обновление железа без остановки сервисов: живая миграция VM освобождает ноду за минуты
- Бэкап с дедупликацией в PBS сокращает окно восстановления VM с суток до десятков минут

ЧТО МЫ ЗАКЛАДЫВАЕМ

Ключевые параметры реализации

9.2

Proxmox VE на Debian 13: ставим ветку 9.x — динамическая CRS-балансировка HA и свежий KVM/QEMU

по докам Proxmox VE 9.2

<5 мс

Потолок задержки сети corosync между нодами; выделенный гигабитный линк только под кластер

по докам pvecm (Cluster Manager)

2 линка

Kronosnet link0+link1: резервный путь кворума через второй NIC — не ловим fence при сбое свитча

наш стандарт

3/2

Пул size=3, min_size=2: три копии, запись живёт при потере ноды, автолечение после возврата

по докам pveceph, Ceph Squid 19.2

60 с

Watchdog (softdog/IPMI): нода без кворума самоизолируется за ~минуту, HA перезапускает её VM

по докам ha-manager

7-4-6

Ретенция vzdump→PBS: keep-daily=7, weekly=4, monthly=6 + еженедельный verify-job чанков

наш стандарт, PBS 4.x

РЕАЛИЗАЦИЯ 1 ИЗ 3

Сборка кластера: сети, кворум, pmxcfs

Что настраиваем

Три ноды (типово Dell/Supermicro, 2 SSD под систему в ZFS RAID1) + управляемый коммутатор с выделенным VLAN под corosync

Как мы это делаем

1. Ставим PVE 9.2 с ISO: root-диск ZFS RAID1 (rpool), сразу переключаем репозиторий на pve-no-subscription либо enterprise по подписке заказчика
2. Разносим сети: vmbr0 — трафик VM, отдельный VLAN corosync (UDP 5405-5412), отдельная сеть 10GbE под storage и миграцию (migration: network в datacenter.cfg)
3. pvecm create <cluster> на первой ноде, pvecm add <IP> --link1 <IP2> на остальных: два kronosnet-линка в corosync.conf с самого начала
4. Проверяем pvecm status: Quorate, expected votes 3; конфиги кластера живут в pmxcfs (/etc/pve) и реплицируются на все ноды автоматически
5. Для площадок из 2 нод добавляем QDevice (corosync-qnetd на внешней машине) — иначе отказ одной ноды кладёт кворум целиком

РЕЗУЛЬТАТ

Кластер управляется из единого веб-интерфейса любой ноды, конфигурация VM синхронна на всех узлах, а потеря линка или свитча не приводит к ложному fence — кворум держится на втором линке.

КЛЮЧЕВОЙ НЮАНС

Corosync — realtime-протокол: он делит сеть с бэкапом или Ceph только до первого шторма трафика. Перед сдачей меряем латентность ping между нодами (<5 мс) и статус обоих линков — corosync-cfgtool -s.

РЕАЛИЗАЦИЯ 2 ИЗ 3

Общее хранилище и HA: Ceph либо ZFS-репликация

Что настраиваем

Гиперконвергентный Ceph на 3 нодах (по 4 SSD/NVMe на OSD) — или ZFS+pvesr для бюджетных площадок из 2-3 нод

Как мы это делаем

1. pveceph install — Ceph Squid 19.2 (в PVE 9.2 доступен и Tentacle 20.2) → по монитору и менеджеру на ноду → pveceph osd create /dev/nvmeXn1 на каждый диск, контролле…
2. Пул VM: size=3, min_size=2, pg_autoscale_mode=on; сеть Ceph public/cluster выносим на 10GbE, отдельно от corosync
3. Бюджетный вариант: локальные ZFS-пулы + pvesr — репликация zvol каждые 15 минут (schedule */15), при аварии теряем максимум один интервал
4. ha-manager add vm:100 --max_restart 3 --max_relocate 2 для критичных VM (1С, SQL, AD); привязка к нодам — ha-manager rules add node-affinity (HA-группы упразднены с…
5. Тестируем при заказчике: выдёргиваем питание ноды — VM с Ceph поднимаются на соседях за 2-3 минуты без ручных действий

РЕЗУЛЬТАТ

Отказ любой из трёх нод не останавливает работу: Ceph продолжает запись на двух копиях, HA-стек перезапускает виртуалки автоматически. Плановое обслуживание — живой миграцией заранее, вообще без даунтайма.

КЛЮЧЕВОЙ НЮАНС

HA имеет смысл только на общем/реплицируемом хранилище. И считаем capacity: кластер обязан жить при N-1 — суммарная нагрузка всех VM должна помещаться на две ноды из трёх.

РЕАЛИЗАЦИЯ 3 ИЗ 3

Перенос VM с VMware и бэкап-контур PBS

Что настраиваем

Миграция парка 20-40 VM со старого гипервизора + отдельный физический сервер (или площадка) под Proxmox Backup Server 4

Как мы это делаем

1. Windows-VM: заранее ставим VirtIO-драйверы (virtio-win ISO) и qemu-guest-agent, затем qm disk import из VMDK/OVF либо штатный мастер импорта ESXi-storage
2. Диски конвертируем в raw на Ceph (qemu-img convert -O raw), контроллер VirtIO SCSI single + discard=on, сеть virtio — без эмуляции e1000/IDE
3. PBS 4 на отдельном железе: datastore на ZFS, подключаем в Datacenter → Storage по fingerprint, шифрование бэкапов ключом клиента
4. Задание vzdump: mode=snapshot (без остановки VM), ежедневно 02:00, ретенция 7-4-6; prune+GC по расписанию на стороне PBS
5. Еженедельный verify-job + тестовое восстановление одной VM в изолированный vmbr — единственное честное доказательство, что бэкап живой

РЕЗУЛЬТАТ

Парк переезжает за 1-2 ночных окна (даунтайм на VM — минуты на финальную синхронизацию), а дедупликация PBS ужимает месяцы ежедневных копий в объём порядка нескольких полных образов. Восстановление VM целиком или отдельного файла — из веб-интерфейса.

КЛЮЧЕВОЙ НЮАНС

Инкрементальность vzdump держится на dirty-bitmap в RAM: после жёсткого рестарта ноды первый бэкап снова прочитает весь диск — закладываем это в окно бэкапа, а не удивляемся утром.

ЧТО МЫ ВЫУЧИЛИ НА ПРАКТИКЕ

Подводные камни

- ✕ **Кластер из 2 нод без QDevice**
  Падение любой ноды рушит кворум, /etc/pve уходит в read-only. Ставим corosync-qnetd на третью машину (хватает и VPS) — голос арбитра решает

- ✕ **Corosync в общей сети**
  Шторм от бэкапа/Ceph поднимает латентность выше 5 мс — ноды теряют токен и самофенсятся. Выделенный VLAN + второй kronosnet-линк обязательны

- ✕ **HA на локальных дисках**
  VM физически нечем поднять на соседней ноде. Для HA — только Ceph/NFS/iSCSI либо ZFS-репликация pvesr с осознанной потерей интервала

- ✕ **Ceph поверх RAID-контроллера**
  Кэш контроллера ломает гарантии записи OSD и добивает латентность. Только HBA/IT-mode, по одному OSD на физический диск, enterprise-SSD с PLP

- ✕ **Бэкап в тот же кластер**
  PBS-датастор на том же Ceph умирает вместе с кластером. PBS — отдельное железо, плюс remote sync копии на вторую площадку

- ✕ **Windows без VirtIO до переноса**
  После импорта VM не видит диск и падает в BSOD. Драйверы virtio-win и qemu-guest-agent ставим ещё на исходном гипервизоре

- ✕ **Обновления без миграции VM**
  Перезагрузка ядра с работающими VM = внеплановый даунтайм. Порядок: живая миграция с ноды → apt dist-upgrade → reboot → возврат VM

- ✕ **Кривое удаление ноды**
  Выключили сервер и забыли — expected votes завышены, кворум хрупкий. Только pvecm delnode, ноду со старым именем/IP назад не возвращаем

УРОВНИ РЕАЛИЗАЦИИ

Как правильно

МИНИМУМ

- 2 ноды PVE 9.x + QDevice-арбитр, ZFS RAID1 под систему на каждой
- ZFS-репликация критичных VM каждые 15 минут (pvesr) между нодами
- PBS на отдельном сервере, ночной vzdump mode=snapshot, ретенция 7-4-6
- Выделенный VLAN под corosync, UPS с NUT-агентом на обеих нодах

НОРМАЛЬНО

- 3 ноды + Ceph Squid (size=3/min_size=2) на enterprise-SSD через HBA
- Два corosync-линка (link0/link1), сеть 10GbE под storage и миграцию
- Node-affinity правила HA для 1C/SQL/AD: max_restart=3, max_relocate=2, тест отказа н...
- Verify-job PBS еженедельно + тестовое восстановление VM раз в месяц

ХОРОШО

- Capacity под N-1: все VM помещаются на две ноды, CRS-балансировка HA
- PBS remote sync на вторую площадку + шифрование датастора ключом клиента
- Мониторинг Zabbix: кворум, статус OSD/PG, latency corosync, задачи vzdump
- Enterprise-подписка Proxmox на прод: обкатанный репозиторий и поддержка вендора

ПРОВЕРЬТЕ У СЕБЯ

Чек-лист самопроверки

- [ ] Переживёт ли инфраструктура отказ одной ноды без ручных действий — проверяли выдёргиванием питания?
- [ ] Есть ли кворум при любой единичной аварии: 3 ноды или 2+QDevice, а не голый двухузловой кластер?
- [ ] Идёт ли corosync по выделенной сети с латентностью до 5 мс и настроен ли второй линк?
- [ ] Хватает ли ресурсов двух нод из трёх, чтобы принять все VM при аварии (правило N-1)?
- [ ] Стоит ли бэкап-сервер (PBS) на отдельном железе, а не внутри защищаемого кластера?
- [ ] Проверяется ли целостность бэкапов verify-job и делали ли тестовое восстановление за последний месяц?
- [ ] Установлен ли qemu-guest-agent во всех VM — консистентные снапшоты и корректный shutdown?
- [ ] Обновляется ли кластер по циклу «миграция VM → обновление → перезагрузка» без остановки сервисов?
- [ ] Закрыт ли веб-интерфейс :8006 от интернета (VPN/ACL) и включена ли 2FA для root@pam?
- [ ] Есть ли план отката и вторая копия бэкапов вне площадки (remote sync PBS)?

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 текущей виртуализации и расчёт кластера: ноды, диски, сети, capacity под N-1
- Развёртывание PVE 9.x под ключ: кластер, Ceph/ZFS, HA-правила, PBS с ретенцией и verify
- Миграция парка VM с VMware/Hyper-V с даунтаймом в минуты на виртуалку, включая 1С и SQL
- Мониторинг кластера в Zabbix и регламентные обновления с живой миграцией без простоя
- Поддержка по SLA: реакция на инциденты, тестовые восстановления, ежеквартальные учения отказа

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 Proxmox VE Administration Guide, гл. Cluster Manager (pvecm) (pve.proxmox.com — 9.2)

02 Proxmox VE Admin Guide, гл. High Availability (ha-manager) (pve.proxmox.com — 9.2)

03 Deploy Hyper-Converged Ceph Cluster (pveceph), Ceph Squid 19.2 (pve.proxmox.com — 9.2)

04 Storage Replication (pvesr) и гл. Backup and Restore (vzdump) (pve.proxmox.com — 9.2)

05 Proxmox Backup Server Documentation: datastore, prune, verify, sync (pbs.proxmox.com — 4.x)

06 Наш шаблон внедрения: чек-лист сборки кластера PVE и учений отказа ноды (itfresh.ru — 2026)

Основано на официальной документации продуктов и нашей практике внедрения.