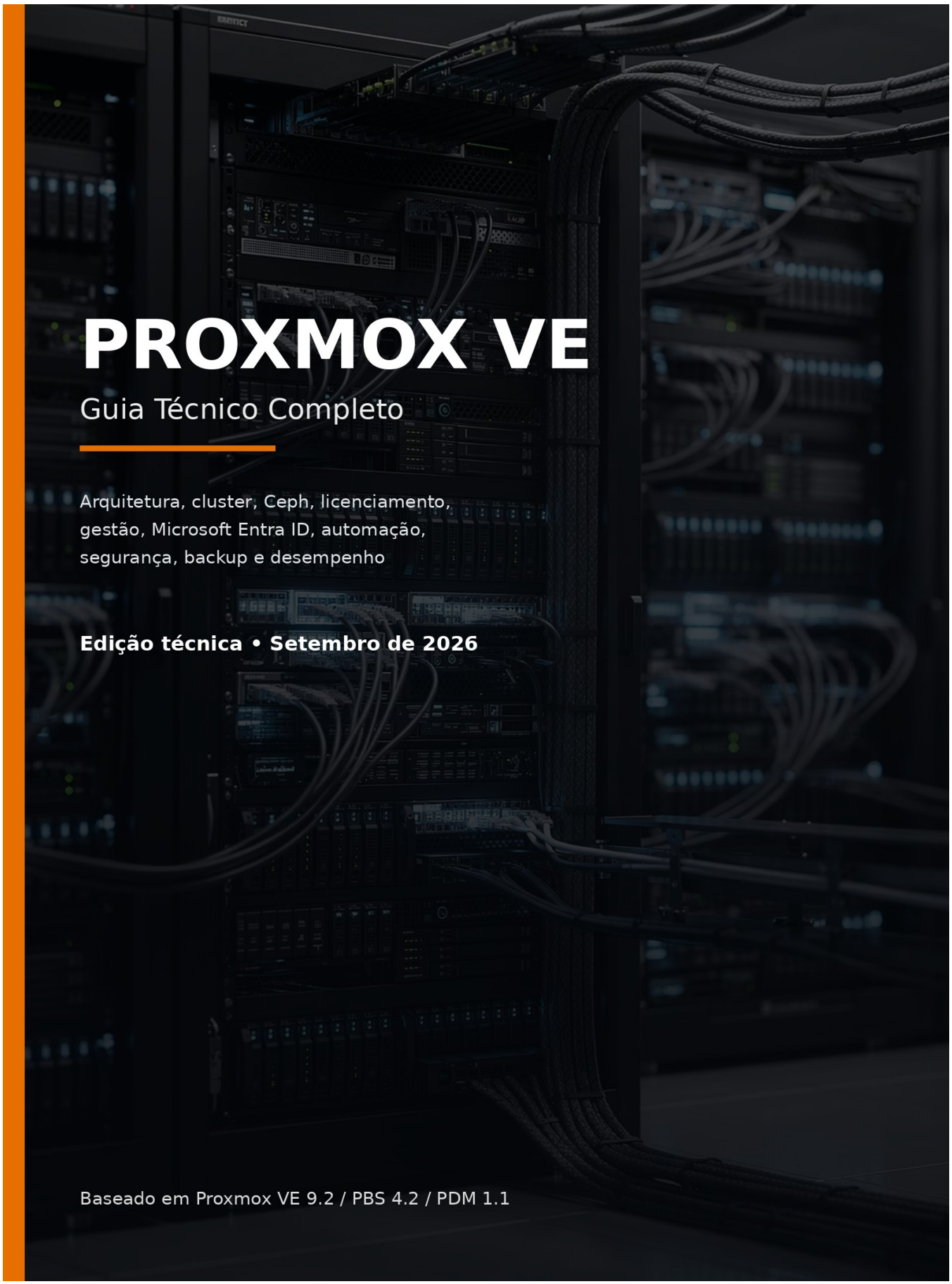

PROXMOX VE

Guia Técnico Completo

Arquitetura, cluster, Ceph, licenciamento, gestão, Microsoft Entra ID, automação, segurança, backup e desempenho

Edição técnica • Setembro de 2026

Baseado em Proxmox VE 9.2 / PBS 4.2 / PDM 1.1

Sobre este e-book

Uma referência prática para desenho, operação e evolução de ambientes Proxmox em produção.

Este material expande a apresentação “Cluster Proxmox: Do Zero ao Avançado” fornecida como base. A estrutura conceitual original - cluster, quorum, Corosync, storage compartilhado, Ceph, migração e HA - foi preservada e aprofundada para um contexto de engenharia de infraestrutura.

A atualização técnica considera o ecossistema disponível em setembro de 2026: Proxmox VE 9.2, Proxmox Backup Server 4.2 e Proxmox Datacenter Manager 1.1. Como repositórios, kernels, QEMU, LXC e Ceph evoluem continuamente, valide sempre as release notes antes de uma implantação ou upgrade.

Como ler os números de desempenho
Resultados publicados são referências de uma configuração específica. Exemplos calculados neste e-book são exercícios de capacidade e não promessas de SLA. CPU, NUMA, controladora, SSD/NVMe, firmware, rede, MTU, filas, workload e política de réplica alteram radicalmente os resultados.

Escopo

- KVM e LXC; arquitetura do host e do cluster; pmxcfs e Corosync; quorum e HA.
- GUI de gestão, Proxmox Datacenter Manager, RBAC e operação multi-site.
- Networking tradicional e SDN; VLAN, VXLAN/EVPN, BGP e WireGuard.
- Storage local e compartilhado; ZFS, NFS, iSCSI, Fibre Channel e Ceph.
- Microsoft Active Directory e Microsoft Entra ID (Azure AD) via OpenID Connect.
- Proxmox Backup Server, retenção, verificação, sync, criptografia e DR.
- REST API, CLI, cloud-init, Ansible, Terraform/OpenTofu e desenho de automação.
- Licenciamento AGPLv3, assinaturas comerciais, custos por socket e suporte.
- Benchmarks publicados, modelos de sizing e três arquiteturas de referência.

Sumário

CAPÍTULO 01

Proxmox em 2026: plataforma e ecossistema

Uma stack aberta que reúne virtualização, containers, storage, rede, backup e gestão distribuída.

Proxmox Virtual Environment (PVE) é uma plataforma de gerenciamento de virtualização baseada em Debian GNU/Linux. Ela combina KVM para máquinas virtuais e LXC para containers Linux em uma única interface, integrando cluster, alta disponibilidade, storage definido por software, rede e disaster recovery. O código do Proxmox VE é publicado sob GNU AGPLv3.

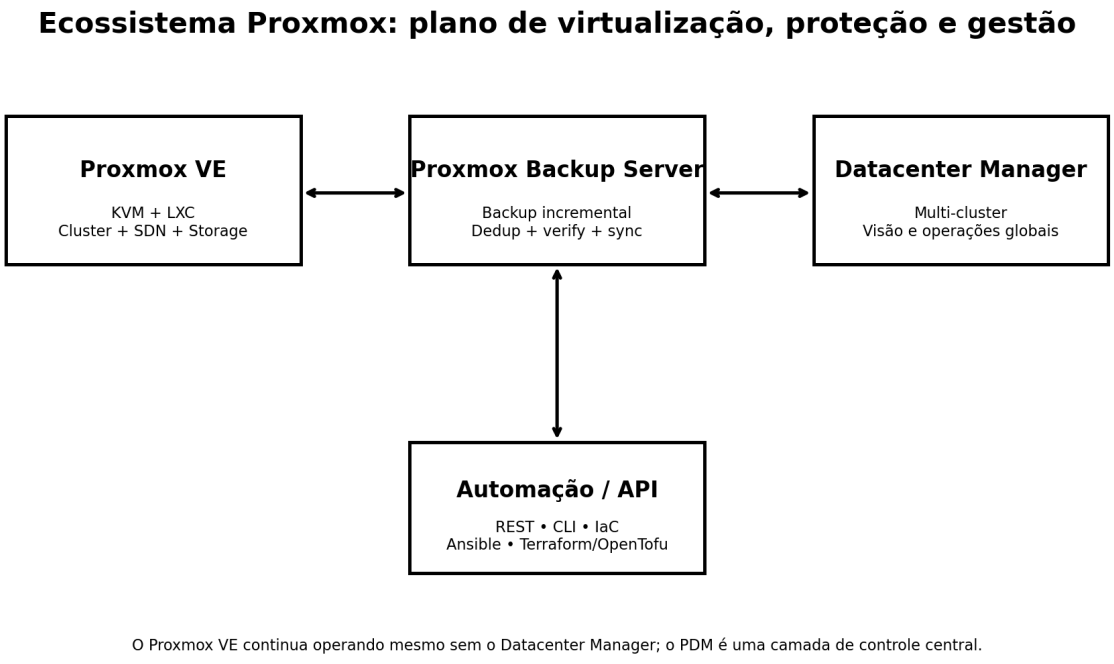

Figura 1 - Visão do ecossistema Proxmox e suas camadas de controle.

Versão de referência

Componente	Versão-base neste guia	Papel
Proxmox VE	9.2	Hypervisor, containers, cluster, HA, SDN e

Componente	Versão-base neste guia	Papel
		storage
Proxmox Backup Server	4.2	Backup incremental, deduplicação, verify, sync e restore
Proxmox Datacenter Manager	1.1	Gestão central de múltiplos clusters/sites
Ceph	Tentacle 20.2 no PVE 9.2	Storage distribuído integrado
Base do PVE 9.2	Debian 13.5 + kernel 7.0	Sistema operacional do host
QEMU / LXC / ZFS	QEMU 11 / LXC 7 / ZFS 2.4	Virtualização, containers e storage local

ARM64 agora é oficial
Em agosto de 2026, a Proxmox lançou uma edição Arm64 do PVE 9.2, com paridade funcional declarada em KVM, LXC, ZFS e Ceph para plataformas oficialmente suportadas como NVIDIA Grace e Vera. Clusters mistos x86-64/Arm64 não devem ser assumidos como suportados para produção sem validação específica.

Onde o Proxmox se posiciona

- Virtualização enterprise: VMs Windows/Linux com KVM/QEMU, VirtIO, UEFI, vTPM e PCIe passthrough.
- Containers de sistema: LXC compartilha o kernel Linux do host e tem overhead de runtime geralmente baixo.
- HCI: compute e Ceph no mesmo conjunto de nodes.
- Infraestrutura tradicional: PVE também funciona com SAN Fibre Channel, iSCSI, NFS e storage de terceiros.
- Multi-site: PDM fornece uma camada de gestão acima de clusters independentes, sem substituir o plano de controle local de cada cluster.

CAPÍTULO 02

Fundamentos: KVM, LXC e arquitetura do host

Entender o que roda no node evita tratar o PVE como uma caixa-preta.

KVM/QEMU

No caminho de VM, KVM fornece virtualização assistida por hardware no kernel Linux e QEMU implementa o modelo de máquina e os dispositivos virtuais. Para workloads modernos, dispositivos VirtIO reduzem overhead de emulação. A escolha de CPU model influencia performance e migração: usar “host” maximiza exposição de recursos da CPU, enquanto modelos compatíveis entre nodes ampliam portabilidade.

LXC

Containers LXC não emulam um kernel completo: processos do container compartilham o kernel do host, isolados por namespaces, cgroups, AppArmor/seccomp e limites de recursos. Isso reduz custo de runtime, mas restringe o guest a Linux e aumenta a importância de hardening do host.

Serviços essenciais do PVE

Serviço / componente	Função operacional
pveproxy	HTTPS/API na porta 8006 e front-end da GUI
pvedaemon	Execução privilegiada de operações do PVE
pvestatd	Coleta periódica de status de VM, storage e nodes
pve-cluster / pmxcfs	Configuração distribuída montada em /etc/pve
corosync	Membership, mensageria e quorum do cluster
pve-ha-crm / pve-ha-lrm	Coordenação e execução local de recursos HA
qemu-server / qm	Gerenciamento de VMs QEMU/KVM
pve-container / pct	Gerenciamento de containers LXC

Princípio de engenharia
Não dependa do node físico como unidade de aplicação. O node deve ser tratado como parte substituível do cluster; identidade e estado persistente devem estar em storage, rede, backup e configuração distribuída.

CAPÍTULO 03

Painel de gestão: GUI e Datacenter Manager

Single pane of glass dentro do cluster; controle global com PDM acima de vários clusters.

GUI nativa do Proxmox VE

A interface web é integrada ao próprio PVE: não é necessário instalar um appliance de gerenciamento separado para operar um cluster. O modelo é multi-master: qualquer node em quorum pode oferecer a visão completa do cluster. Isso remove o “manager node” dedicado como ponto único de administração.

- Árvore de recursos: Datacenter → cluster/nodes → VMs/CTs → storages.
- Visão Datacenter: cluster, HA, opções globais, jobs, permissões, realms, firewall e SDN.
- Visão Node: shell, system, network, disks, ZFS, Ceph, updates, logs e métricas.
- Visão VM/CT: hardware, options, console, snapshots, backup, firewall, replication e métricas.
- Task log: cada operação gera um UPID rastreável, útil para troubleshooting e auditoria.

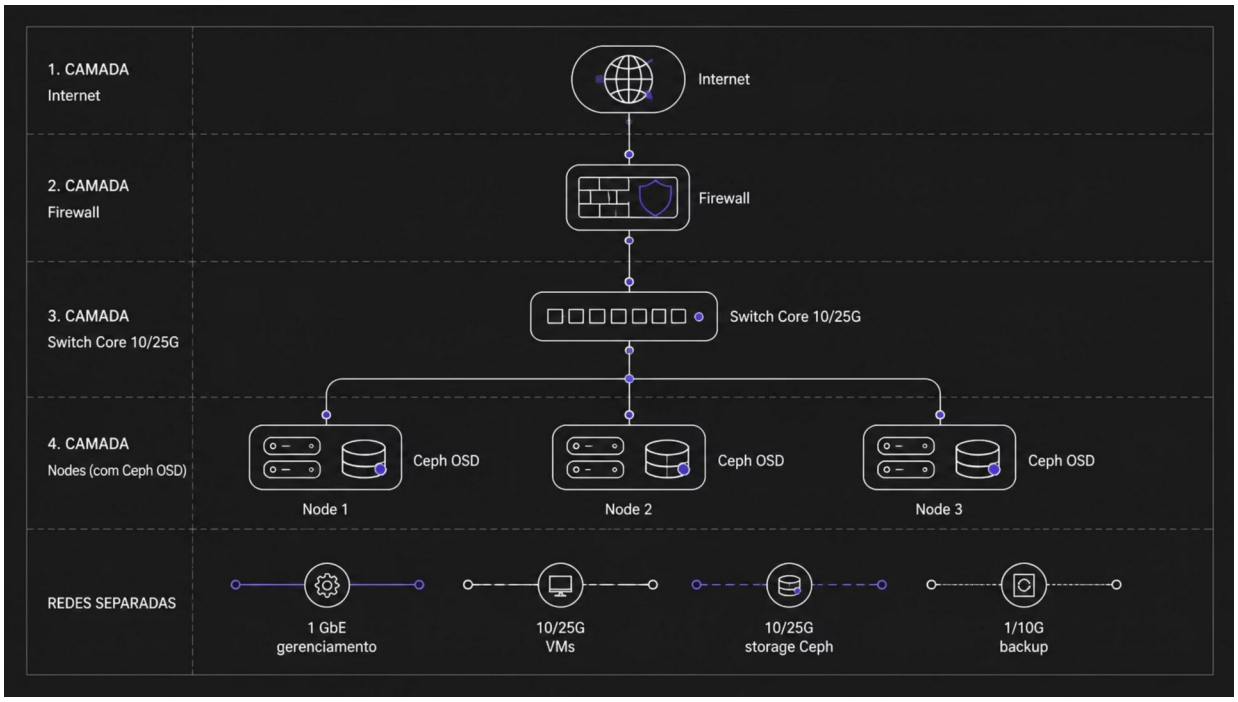

Imagem real da apresentação-base: exemplo de visualização de arquitetura/topologia usada para explicar o ambiente.

Proxmox Datacenter Manager 1.1

O PDM foi criado para ambientes distribuídos e grandes frotas. Ele conecta “remotes” - clusters PVE, nodes standalone e instâncias PBS - e consolida inventário, métricas, status, tarefas, atualizações e operações de ciclo de vida. A arquitetura é desacoplada: a indisponibilidade do PDM não interrompe as VMs dos clusters gerenciados.

Capacidade PDM 1.1	Uso
Dashboard global	Saúde, CPU, RAM, storage, remotes e visão geográfica
Guest management	Lista VMs/CTs entre remotes, start/stop/shutdown/resume
Snapshots	Criar, editar descrição, rollback e remover em visão central
Ceph monitoring	Saúde e performance de OSD, MON, MGR, MDS, pools e CephFS
Subscription registry	Pool central de chaves e associação a remotes
Automated installation	Answer files, tokens e acompanhamento da instalação
Cross-cluster migration	Movimentação de guests entre clusters independentes
Updates	Visibilidade e execução central de atualizações

PDM não substitui o PVE
Configurações avançadas ainda podem abrir a GUI nativa do remote. Trate o PDM como plano de gestão global, não como requisito de funcionamento das VMs.

CAPÍTULO 04

Cluster, pmxcfs, Corosync e quorum

O cluster é primeiro uma estrutura de consenso e configuração; HA é uma função adicional.

No material-base, o cluster é corretamente apresentado como união de nodes sob um Datacenter lógico, com gestão central, migração e base para HA/Ceph. O ponto crítico permanece: cluster não significa HA automaticamente. Para reinício automático de guests após falha, é necessário combinar quorum saudável, storage acessível e recursos registrados no HA.

pmxcfs: configuração distribuída

O Proxmox Cluster File System (pmxcfs) é um filesystem em userspace montado em /etc/pve. Ele replica configurações em tempo real entre nodes via Corosync, faz checagens de consistência e torna-se read-only quando o node perde quorum. A configuração de VMs, storages, usuários, firewall, HA e SDN vive nesse plano de controle.

Quorum

Topologia	Votos úteis	Falha tolerável antes de perder maioria	Comentário
2 nodes	2	0 sem qdevice	Use qdevice para terceiro voto em produção
3 nodes	3	1	Padrão mínimo clássico para HA
5 nodes	5	2	Maior margem de falha
7 nodes	7	3	Comum em clusters maiores; ainda exige desenho de rede

Quorum protege contra split-brain. Se uma partição de rede deixar subconjuntos do cluster sem maioria, esses lados não devem continuar tomando decisões conflitantes sobre os mesmos recursos.

Corosync

- Use rede estável, previsível e de baixa latência; jitter e perda são tão importantes quanto banda.
- Separe tráfego de storage e de VM do tráfego de cluster sempre que possível.
- Implemente redundância de links Corosync/Knet para remover uma NIC/switch como ponto único de falha.
- Não estique um único cluster por WAN de alta latência sem uma análise específica de failure domains e quorum.

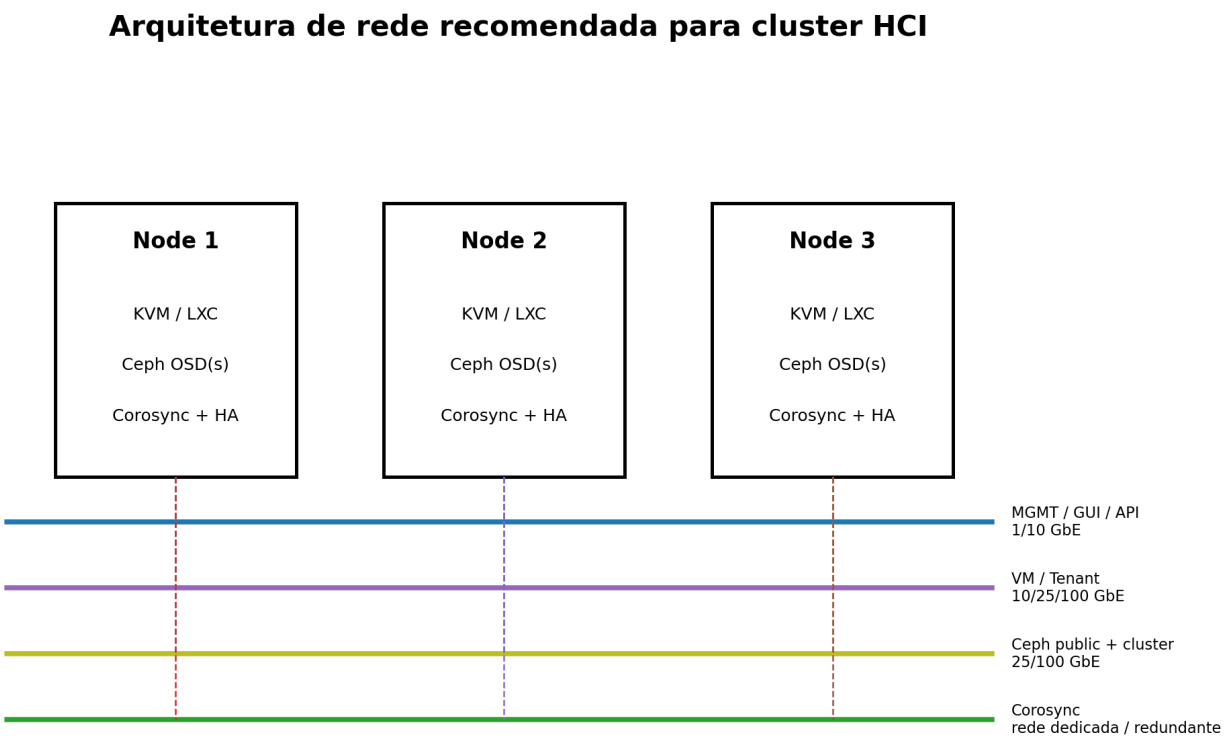

Figura 2 - Segmentação lógica de redes em um cluster hiperconvergente.

CAPÍTULO 05

Alta disponibilidade e Dynamic Load Balancer

Detecção de falha, fencing, restart e balanceamento são mecanismos distintos que trabalham juntos.

O que o HA realmente entrega

Quando um node falha, o HA tenta recuperar os serviços configurados iniciando-os em nodes elegíveis. Isso não é “zero downtime”: existe um intervalo para detecção, fencing/lock e boot do guest. Zero downtime no nível de aplicação exige mecanismos dentro do software, como banco replicado, cluster de aplicação ou balanceador com múltiplas instâncias.

Componentes internos

Elemento	Responsabilidade
CRM - Cluster Resource Manager	Decide o estado desejado e coordena recursos HA no cluster
LRM - Local Resource Manager	Executa start/stop/migrate no node local
Watchdog / fencing	Evita dois nodes executando o mesmo recurso em cenários de falha
HA rules/groups	Restringem/ordenam posicionamento e afinidades
Maintenance workflow	Prepara node para manutenção sem perder estado operacional

Dynamic Load Balancer no PVE 9.2

O PVE 9.2 adiciona um Dynamic Load Balancer ligado ao Cluster Resource Scheduler (CRS). O objetivo é melhorar o posicionamento de workloads e rebalancear guests conforme utilização e regras, reduzindo operações manuais em clusters maiores. Isso é balanceamento de infraestrutura; não substitui autoscaling de aplicação.

Capacidade N-1
Se o cluster roda normalmente a 85-90% de RAM ou CPU, HA pode existir “no papel” e falhar

na prática por falta de capacidade de destino. Sizing de produção deve reservar capacidade para a maior falha planejada.

CAPÍTULO 06

Rede: bridges, VLAN, SDN, EVPN/BGP e WireGuard

Da bridge Linux simples até malhas multi-tenant e roteamento distribuído.

Linux Bridge

A Linux Bridge é o caminho mais direto: NIC física → bridge vmbrX → interfaces VirtIO das VMs. Com VLAN-aware, uma bridge trunk pode transportar várias VLANs e cada guest recebe uma tag. É simples, eficiente e adequado a muitos ambientes enterprise.

```
auto vmbr0
iface vmbr0 inet static
    address 10.0.0.11/24
    gateway 10.0.0.1
    bridge-ports bond0
    bridge-stp off
    bridge-fd 0
    bridge-vlan-aware yes
```

Bonding

- active-backup: redundância simples, sem depender de LACP.
- 802.3ad/LACP: agrega links e distribui fluxos, desde que o switch esteja corretamente configurado.
- Evite assumir que um único fluxo TCP usará a soma de todos os links; hashing normalmente mantém o fluxo em um membro do bond.

SDN

O SDN do PVE permite criar zonas e VNets de forma central. A stack evoluiu para VXLAN/EVPN, BGP e, no 9.2, fabrics WireGuard e políticas de roteamento mais granulares. O uso cresce em multi-tenant, edge e ambientes onde a rede virtual precisa acompanhar o lifecycle dos workloads.

Tecnologia	Use quando	Ponto de atenção
VLAN	Segmentação L2 tradicional	Escala limitada pelo domínio L2 e IDs de VLAN

Tecnologia	Use quando	Ponto de atenção
VXLAN	Overlays e multi-tenant	Exige underlay IP consistente e MTU coerente
EVPN/BGP	Controle distribuído de reachability	Requer domínio de roteamento bem operado
WireGuard Fabric	Conectividade segura entre endpoints/sites	Planejar MTU, chaves e rotas
Firewall distribuído	Política próxima ao guest	Regras e objetos precisam de governança

CAPÍTULO 07

Storage: ZFS, NFS, SAN/iSCSI/FC e plugins

O storage determina grande parte da disponibilidade, performance e complexidade operacional.

Storage local

Local-LVM, LVM-thin e ZFS entregam baixa latência e simplicidade por host. A contrapartida é a mobilidade: live migration de uma VM com disco local precisa copiar o storage, e HA depende de replicação ou outro mecanismo para o destino ter os dados.

ZFS

- Checksums end-to-end, snapshots, compressão, ARC e RAID-Z/mirror.
- Para VM random I/O, mirrors tendem a entregar perfil de IOPS superior a RAID-Z de mesma quantidade de discos.
- Planeje RAM para ARC e não use hardware RAID por baixo do ZFS quando o objetivo é que o ZFS controle os discos.
- ZFS replication no PVE é útil para DR/HA com RPO periódico, mas não é a mesma coisa que storage síncrono compartilhado.

Storage compartilhado

Backend	Modelo	Pontos fortes	Riscos / limites
NFS	File	Simples, amplo suporte, snapshots conforme NAS	Servidor/array precisa ser HA e bem dimensionado
iSCSI + LVM	Block	SAN tradicional, multipath, performance previsível	Operação de LUN/paths e dependência do array
Fibre Channel	Block	Baixa latência, zoning e SAN madura	Custo/complexidade de fabric e multipath
Ceph RBD	Distribuído	Sem array central, self-healing, scale-	Rede e operação são críticas

Backend	Modelo	Pontos fortes	Riscos / limites
		out	
ZFS local	Local	Latência, snapshots, independência de SAN	Migração de disco e HA exigem estratégia adicional

Shared storage e live migration
Quando todos os nodes acessam o mesmo volume, a migração normal não precisa transferir o disco da VM; principalmente memória e estado de CPU/dispositivos são migrados. Isso reduz drasticamente o volume de dados no caminho de migration.

CAPÍTULO 08

Ceph hiperconvergente em profundidade

Ceph elimina o array central, mas troca simplicidade por distribuição, consenso e tráfego de réplica.

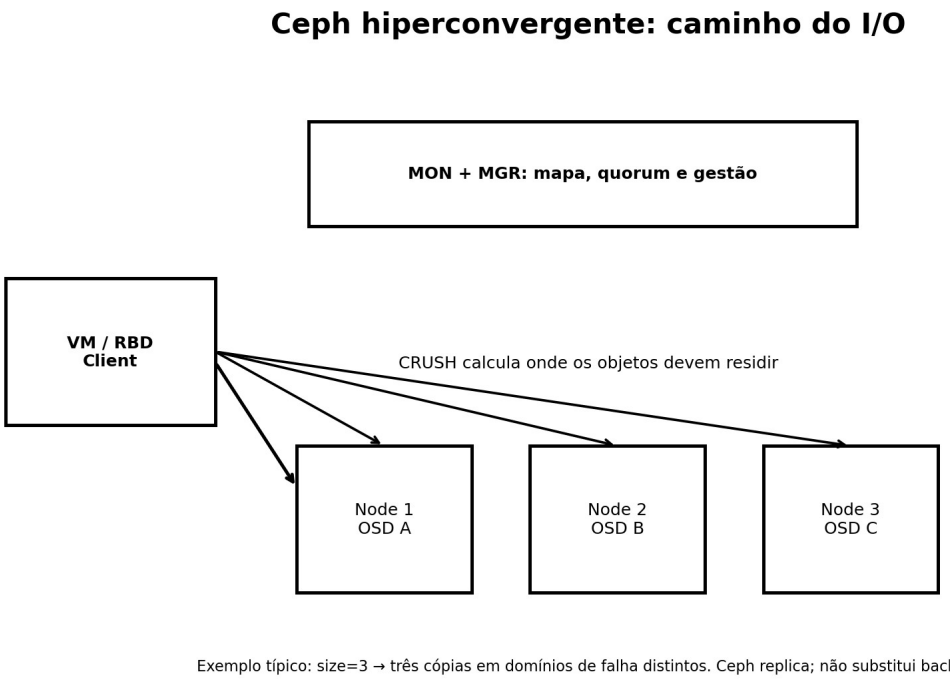

Figura 3 - Componentes básicos e caminho de I/O do Ceph integrado ao Proxmox.

Componentes

Componente	Função
MON	Mantém mapas e quorum do cluster Ceph
MGR	Métricas, módulos de gestão e interfaces auxiliares
OSD	Armazena objetos e participa de réplica/recovery

Componente	Função
CRUSH	Algoritmo que determina placement por regras/failure domains
RBD	Dispositivo de bloco para discos de VMs
CephFS	Filesystem POSIX distribuído, útil para conteúdos de arquivo

Capacidade e réplica

Em um pool replicado com size=3, um dado lógico é armazenado em três OSDs/failure domains. Portanto, uma aproximação inicial de capacidade lógica é raw/3, antes de reservas, metadata e headroom de operação.

Exemplo	Cálculo	Resultado
3 nodes × 4 NVMe × 3,84 TB	3 × 4 × 3,84	46,08 TB raw
Pool size=3	46,08 / 3	15,36 TB lógicos nominais
Meta operacional 75%	15,36 × 0,75	11,52 TB de ocupação-alvo

Rede Ceph

O benchmark oficial Proxmox de 2023/12 mostrou que 10 Gb/s pode saturar com apenas um SSD muito rápido; 25 Gb/s também pode virar gargalo; 100 Gb/s deslocou o gargalo em parte para o cliente Ceph no ambiente testado. Em 2026, isso reforça 25/100 GbE como desenho natural para HCI NVMe, dependendo do número de OSDs e workload.

Rebuild e recovery

A falha de um OSD/node não só reduz redundância: dispara tráfego de recuperação. O cluster precisa ter banda e IOPS livres para servir aplicações e recuperar réplicas ao mesmo tempo. Por isso, performance em estado degradado deve fazer parte do teste de aceite.

Ceph não é backup
Replicação protege contra falha física. Exclusão acidental, ransomware ou corrupção lógica também são replicados. Use PBS e cópias fora do cluster.

CAPÍTULO 09

Migração de VMs e compatibilidade de CPU

A migração é uma operação de memória, CPU, dispositivos, rede e, às vezes, disco.

Offline x live migration

Na migração offline o guest é desligado e movido. Na live migration, o guest continua executando enquanto páginas de memória são copiadas e reenviadas conforme ficam “sujas”. O switchover final é curto quando a taxa de dirty pages converge abaixo da capacidade de transferência.

CPU model

- “host”: maior exposição de features e potencial de performance; menor portabilidade entre gerações diferentes.
- Modelo comum/custom CPU: define um baseline para nodes heterogêneos e facilita live migration.
- Antes de expansão de cluster, valide flags de CPU, microcode e firmware - não apenas fabricante e número de cores.

Exemplo calculado: memória

Uma VM com 64 GiB migrando por uma rede 25 GbE com throughput efetivo de 2,7 GB/s teria um primeiro passe de aproximadamente 24 segundos (64/2,7). O tempo real será maior se a VM sujar memória rapidamente, se houver congestionamento ou se a criptografia/CPU limitar o fluxo.

Exemplo calculado: disco local

Transferir 1 TiB de disco por 25 GbE a 2,5 GB/s efetivos tem piso teórico em torno de 7 minutos; metadados, cache, storage de origem/destino e I/O concorrente elevam o tempo. Com storage compartilhado, esse volume de disco normalmente deixa de fazer parte da migração.

CAPÍTULO 10

Backup, restore e Proxmox Backup Server

Alta disponibilidade reduz indisponibilidade; backup reduz perda de dados e amplia opções de recuperação.

PBS 4.2

O Proxmox Backup Server recebe backups de VMs, containers e hosts, envia dados de forma incremental e deduplica chunks no datastore. Cada snapshot continua representando um backup completo do ponto de vista de restauração. O PBS usa checksums SHA-256 para verificação de integridade e suporta criptografia client-side.

Função	Impacto operacional
Incremental + deduplicação	Reduz rede e capacidade consumida entre pontos de backup
Verify jobs	Detecta corrupção com checksums e leitura periódica
Prune + GC	Aplica retenção e remove chunks não referenciados
Sync jobs	Replica datastores/namespace para outro PBS
Client-side encryption	Servidor remoto pode armazenar dados sem conhecer conteúdo
Tape	Camada offline/air-gap física para retenção longa
S3-backed storage (4.2)	Expande opções de backend para backup

Regra 3-2-1-1-0 aplicada

- 3 cópias dos dados, contando produção.
- 2 tipos/locais de mídia ou failure domains.
- 1 cópia off-site.
- 1 cópia offline, imutável ou fortemente isolada.
- 0 erros não detectados: jobs de verify e testes de restore.

Sizing rápido do PBS

A documentação do PBS recomenda, para produção, CPU moderna com pelo menos 4 cores, mínimo de 4 GiB para sistema/cache/daemons e aproximadamente 1 GiB adicional de RAM por TiB de storage. Esse número é um ponto de partida: dedup, verify, filesystem, paralelismo e janela de backup podem exigir mais memória e IOPS.

RPO e RTO são decisões de negócio
Backups a cada 24 horas não entregam RPO de 15 minutos. Da mesma forma, ter backup não garante RTO baixo se restore de centenas de TB nunca foi ensaiado.

CAPÍTULO 11

Identidade: AD, LDAP e Microsoft Entra ID

Autenticação central reduz contas locais; RBAC continua sendo a camada de autorização do Proxmox.

Realms suportados

Realm	Use típico	Observação
pam	Break-glass / administração Linux	Conta do sistema; preserve acesso de emergência
pve	Usuários internos do PVE	Simples e totalmente local
LDAP	Diretórios LDAP genéricos	Pode sincronizar usuários/grupos
Active Directory	AD on-prem	Integração nativa do diretório Microsoft
OpenID Connect	SSO moderno / Entra / Keycloak / Google	Autenticação federada via browser

Active Directory on-premises

O realm AD é adequado quando o diretório corporativo está acessível via LDAP/LDAPS e a organização quer sincronização de identidades e grupos. A autorização continua em ACLs do PVE: por exemplo, grupo de operadores com PVEVMAdmin em um Resource Pool e auditores com PVEAuditor no Datacenter.

Microsoft Entra ID (Azure AD) via OIDC

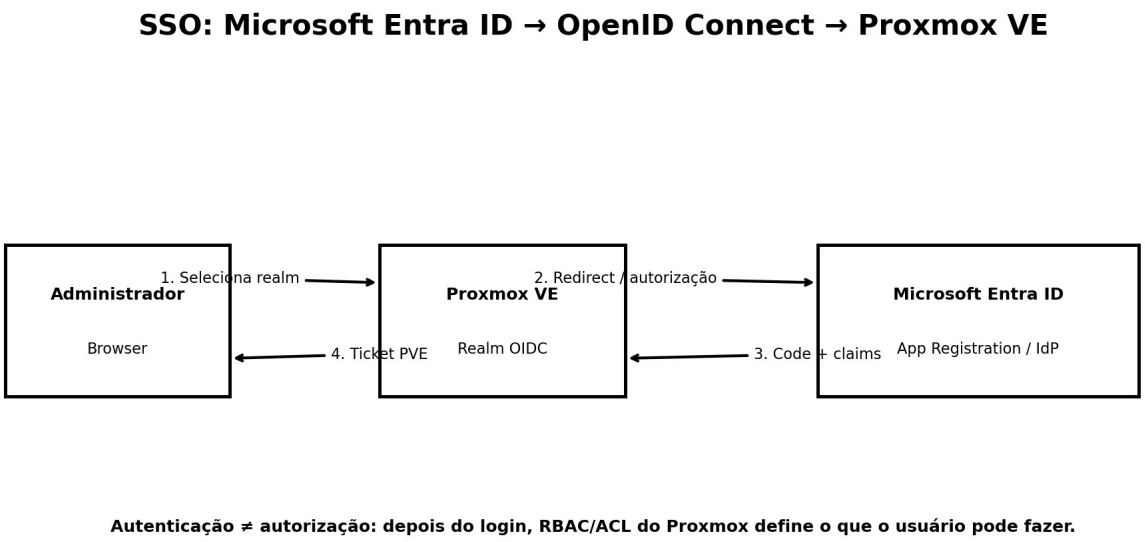

Figura 4 - Fluxo conceitual de SSO usando Microsoft Entra ID e OpenID Connect.

Para Entra ID, o desenho moderno é OpenID Connect. O PVE usa discovery a partir do issuer URL. Em um tenant Microsoft, o issuer normalmente segue o padrão https://login.microsoftonline.com/<tenant-id>/v2.0. Registre o PVE como aplicação Web/confidential client, configure Client ID e Client Secret e cadastre exatamente os redirect URIs usados pelos administradores, como https://pve1.exemplo.com:8006.

```
pveum realm add entra \
  --type openid \
  --issuer-url https://login.microsoftonline.com/<TENANT-ID>/v2.0 \
  --client-id <CLIENT-ID> \
  --client-key <CLIENT-SECRET> \
  --username-claim email \
  --autocreate 1
```

Autocreate não concede privilégios
Criar automaticamente o usuário após o primeiro login não significa torná-lo administrador. Defina grupos, roles e ACLs de menor privilégio. Mantenha um usuário local de emergência e teste o login OIDC antes de torná-lo padrão.

Boas práticas Entra ID

- App registration dedicada ao ambiente Proxmox; não reutilize segredo de outra aplicação.
- DNS e certificado TLS válidos para a URL de administração.
- MFA/Conditional Access no Entra para administradores, conforme política da empresa.

- Rotação de client secret e monitoramento de expiração.
- Grupo de segurança dedicado para acesso; mapeamento e ACLs documentados.
- Teste de falha do IdP: operador deve continuar capaz de usar uma conta break-glass local.

CAPÍTULO 12

Segurança: RBAC, MFA, firewall, API tokens e hardening

O hypervisor é uma camada privilegiada: reduzir superfície e privilégios é obrigatório.

RBAC e ACL

No Proxmox, permissões combinam sujeito (user/group/token), role (conjunto de privilégios) e path (objeto). Essa estrutura permite separar plataforma, storage, redes, backup, tenants e auditoria. Resource Pools ajudam a delegar conjuntos de VMs e storages sem entregar acesso ao cluster inteiro.

MFA

- TOTP e WebAuthn são mecanismos comuns para contas locais/realms suportados.
- Para OIDC, MFA normalmente é realizado no Identity Provider, como Microsoft Entra ID.
- Break-glass deve ser protegido com credencial forte, MFA quando possível e acesso administrativo controlado.

API tokens

Use tokens para automação em vez de senhas humanas. Tokens podem ter privilégios separados do usuário e devem receber apenas permissões necessárias. Guarde segredos em vault, não em scripts ou repositórios Git.

Firewall

O firewall distribuído do PVE aplica regras nos nodes e pode operar nos níveis Datacenter, Node e VM/CT. Security Groups, IP Sets, aliases e macros reduzem duplicação. Em multi-tenant, combine firewall, VLAN/VXLAN e RBAC; nenhum deles isoladamente representa segmentação completa.

Checklist de hardening

- Rede de gerenciamento fora da Internet pública; acesso via VPN/bastion.
- TLS com certificado confiável e DNS consistente.

- SSH com chaves, root remoto restrito conforme política, e logging central.
- Enterprise repository em produção; janela e processo de patching.
- MFA para administradores e revisão periódica de ACLs/API tokens.
- NTP confiável - horário afeta cluster, logs, certificados e autenticação.
- Firmware/BMC isolados em rede de gestão e com credenciais próprias.
- Backup da configuração e testes de recuperação do cluster.

CAPÍTULO 13

Orquestração e Infrastructure as Code

Proxmox oferece primitives; a automação transforma primitives em serviço repetível.

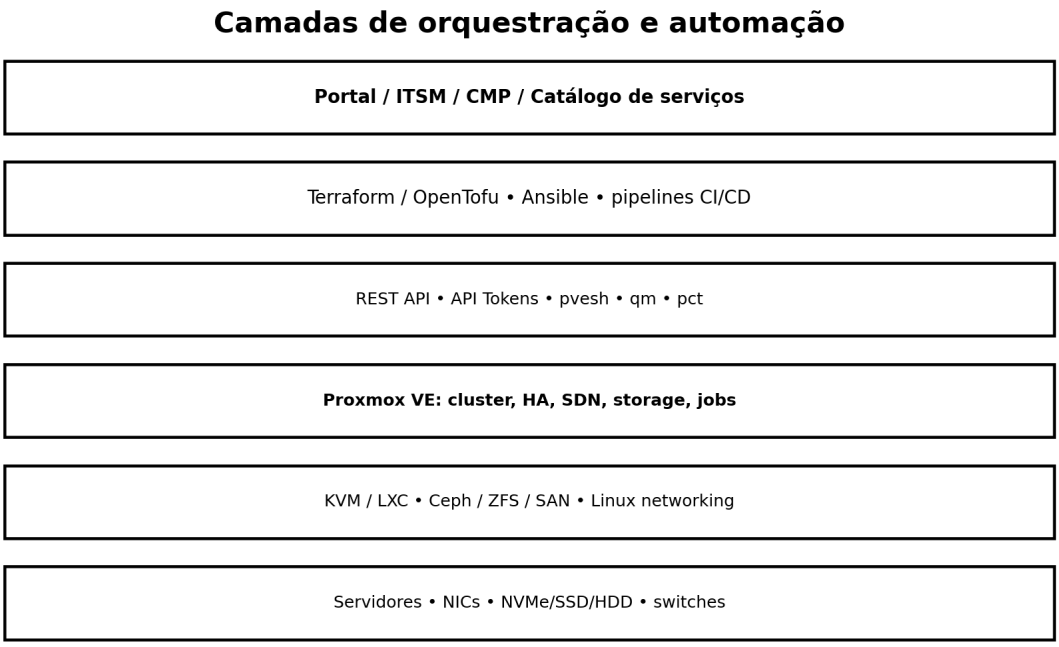

Figura 5 - Camadas de orquestração, do hardware ao portal de autosserviço.

REST API

A API REST do PVE expõe praticamente as mesmas operações da GUI. O modelo usa JSON e schema formal, o que facilita integração com portais, ITSM, CMDB e pipelines. A GUI é, em grande parte, um cliente dessa API.

```
curl -k -H 'Authorization: PVEAPIToken=<USER>@<REALM>!<TOKENID>=<SECRET>' \
  https://pve1.exemplo.com:8006/api2/json/cluster/resources
```

CLI

Ferramenta	Uso
qm	Lifecycle de VMs: create, set, clone, migrate, snapshot
pct	Lifecycle de containers LXC
pvesh	Chamada local da árvore de API
pvesm	Storage manager
pvecm	Cluster e Corosync
pveum	Usuários, grupos, realms, ACLs e tokens
ha-manager	Recursos e regras HA

Cloud-init e templates

```
qm clone 9000 120 --name app-prod-01 --full 1
qm set 120 --cores 8 --memory 16384
qm set 120 --ipconfig0 ip=10.20.30.51/24,gw=10.20.30.1
qm set 120 --sshkeys /root/keys/platform.pub
qm start 120
```

Ansible / Terraform / OpenTofu

A própria Proxmox posiciona Ansible e Terraform/OpenTofu como caminhos de automação. Em projetos reais, trate providers/collections como componentes de software com versão fixa, testes e pipeline. Não misture mudanças manuais e IaC sem uma política clara de ownership, porque drift é inevitável.

O que “orquestração” significa no PVE

- Cluster Resource Scheduler: posicionamento/balanceamento de infraestrutura.
- HA Manager: recovery e manutenção de guests marcados como HA.
- Jobs: backup, replication, verify/sync no PBS, updates e tarefas agendadas.
- PDM: operações cross-cluster e centralização multi-site.
- IaC externo: criação declarativa de VMs, redes, templates, tags, pools e integração de service catalog.

PVE não é Kubernetes
PVE orquestra infraestrutura virtual. Kubernetes orquestra aplicações containerizadas. Os dois podem coexistir: VMs no PVE hospedam control planes/workers Kubernetes, enquanto

PVE cuida de lifecycle, HA, rede e storage da infraestrutura.

CAPÍTULO 14

Observabilidade, logs, métricas e operação

Um cluster saudável precisa ser observável antes da primeira falha.

O que acompanhar

Domínio	Indicadores-chave
Node	CPU, load, steal/pressure, RAM, swap, temperatura, filesystem, kernel errors
Cluster	Quorum, links Corosync, tasks falhas, HA state, versões divergentes
VM/CT	CPU, RAM, ballooning, network, disk throughput/latency
Ceph	HEALTH, OSD up/in, PG states, latency, recovery/backfill, fullness
Network	drops/errors, interface utilization, LACP, MTU mismatch, retransmits
PBS	job duration, dedup ratio, verify errors, GC, datastore fullness, restore tests

External metrics

O PVE pode enviar métricas para sistemas externos e o ecossistema tem integração ampla com soluções como Zabbix, Prometheus/InfluxDB via exporters e Grafana. A plataforma também mantém histórico RRD para gráficos operacionais. Para enterprise, correlacione métricas do PVE com switch, BMC, storage e aplicação.

SLO operacional sugerido

- Quorum: 100% dos períodos de produção, excluindo manutenção planejada.
- Ceph: HEALTH_OK como estado esperado; warning deve gerar ticket e investigação.
- Backup: sucesso por job + RPO medido, não apenas “job executou”.
- Restore: teste automatizado/periódico de amostra; RTO medido.
- Capacidade: alertar antes de 70-75% de RAM/storage em clusters sem grande folga.

CAPÍTULO 15

Licenciamento e assinatura comercial

Software livre não significa “sem governança”: separe licença de uso, repositório e suporte.

Licença do software

O Proxmox VE é software livre sob GNU AGPLv3. Não existe uma edição “capada” sem subscription: recursos de HA, live migration, cluster, backup, storage e networking estão disponíveis independentemente do nível de assinatura. A assinatura comercial diferencia principalmente repositório enterprise, suporte e SLA.

Modelo de assinatura - setembro de 2026

Plano	€/socket/ano	Tickets / resposta publicada	Uso típico
Community	120	Sem tickets de suporte	Lab, treinamento e não crítico
Basic	370	3 / 1 dia útil	Produção menor, equipe autônoma
Standard	550	10 / 4 horas*	Produção enterprise com escalonamento
Premium	1.100	Ilimitados / 2 horas*	Missão crítica e maior cobertura

* Resposta é primeiro retorno para casos críticos, conforme condições do suporte. Valores líquidos, por socket físico ocupado, por ano, sujeitos a alteração e impostos.

Regras importantes

- Conta-se socket físico ocupado; quantidade de cores não altera o preço.
- Todos os nodes de um mesmo cluster devem manter o mesmo nível de subscription.
- A subscription é anual.
- Basic/Standard/Premium incluem o Enterprise Repository e suporte comercial.

- PDM é open source; para suporte enterprise do PDM, remotes com assinaturas ativas qualificadas dão cobertura sem uma chave separada do PDM.

Exemplos de custo anual

Cluster	Sockets ocupados	Community	Basic	Standard	Premium
3 nodes × 1 socket	3	€ 360	€ 1.110	€ 1.650	€ 3.300
3 nodes × 2 sockets	6	€ 720	€ 2.220	€ 3.300	€ 6.600
5 nodes × 2 sockets	10	€ 1.200	€ 3.700	€ 5.500	€ 11.000

Mudança anunciada em 02/09/2026
A Proxmox anunciou suporte enterprise 24/7 global a partir de 19/10/2026. Premium terá 24/7 desde o início, Standard entra em uma janela de onboarding no Q4/2026 e Basic mantém o SLA existente. Na data desta edição, a mudança foi anunciada, mas ainda não estava vigente.

CAPÍTULO 16

Desempenho: benchmark, rede, storage e latência

Desempenho em virtualização é uma cadeia: guest → hypervisor → kernel → rede/storage → hardware.

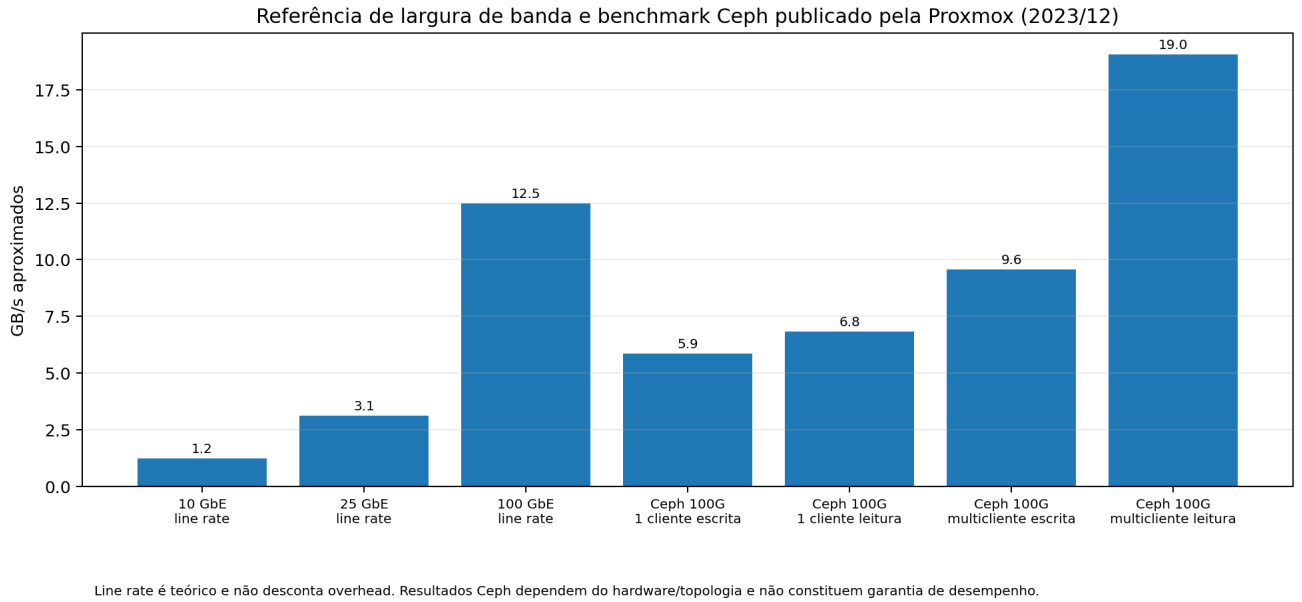

Figura 6 - Comparação entre line rate teórico e resultados agregados do benchmark Ceph oficial Proxmox 2023/12.

Benchmark Ceph publicado pela Proxmox

No benchmark 2023/12 de um cluster HCI com SSDs rápidos, a Proxmox reportou: 10 Gb/s facilmente saturado; 25 Gb/s também podendo se tornar gargalo; em 100 Gb/s, um cliente alcançou até aproximadamente 6.000 MiB/s de escrita e 7.000 MiB/s de leitura. Com múltiplos clientes em paralelo, foram observados até aproximadamente 9.800 MiB/s de escrita e 19.500 MiB/s de leitura.

Não extrapole benchmark
Esses números são resultados do laboratório publicado, não uma especificação do produto. Uma VM individual pode ficar muito abaixo do throughput agregado de RADOS/Ceph por limites de fila, iodepth, CPU, RBD, filesystem, bloco e aplicação.

Line rate de rede

Link	Gb/s	Máximo decimal teórico	Máximo binário aproximado
1 GbE	1	125 MB/s	119 MiB/s
10 GbE	10	1,25 GB/s	1,16 GiB/s
25 GbE	25	3,125 GB/s	2,91 GiB/s
40 GbE	40	5,0 GB/s	4,66 GiB/s
100 GbE	100	12,5 GB/s	11,64 GiB/s

Perfis de workload

Workload	Priorize	Teste
Banco OLTP	latência p99, fsync, 4K random	fio randrw + benchmark do banco
VDI	IOPS/latência em boot storm	4K/8K random, muitos jobs
Backup	throughput sequencial + dedup CPU	janela real e restore
AI / dados	throughput, NIC, GPU passthrough/vGPU	pipeline real, não só fio
Web/app geral	CPU scheduler, RAM e latência moderada	stress-ng + teste de aplicação

CAPÍTULO 17

Sizing: CPU, RAM, storage, rede e reserva N-1

Dimensionar pelo pico e pela falha, não pela média diária.

CPU

Não existe um ratio universal de vCPU:pCPU. Como regra de engenharia - não limite do PVE - workloads críticos e sensíveis à latência podem ficar próximos de 1:1 a 2:1; servidores gerais frequentemente aceitam oversubscription maior; dev/test pode ir além. O indicador decisivo é latência do workload, utilização e contenção observada.

Memória

Memória de guest é mais rígida que CPU: overcommit excessivo força ballooning/swap e produz degradação abrupta. Reserve RAM para o host, cache/storage e, em HCI, para Ceph. Em clusters HA, calcule o node mais carregado falhando e redistribua suas VMs nos sobreviventes.

Exemplo N-1

Item	3 nodes	Após falha de 1 node
RAM física	3 × 512 GiB = 1.536 GiB	1.024 GiB disponíveis
Reserva host/Ceph	~64 GiB por node (exemplo)	~128 GiB nos 2 restantes
RAM útil aproximada	1.344 GiB no estado normal	~896 GiB após falha
Meta de guests para N-1	≤ ~896 GiB	Cabe nos 2 sobreviventes

O número de 64 GiB reservado por node é apenas um exemplo de projeto. Em Ceph com muitos OSDs/NVMe, o consumo real deve ser medido e dimensionado conforme configuração.

Storage

- Capacity: raw × eficiência de proteção × headroom.

- Performance: IOPS e throughput por device, multiplicados de forma realista pela topologia RAID/Ceph.
- Latency: medir p95/p99 sob carga e em recovery, não só average.
- Endurance: DWPD/TBW de SSDs enterprise, especialmente para Ceph e bancos.
- Failure domain: chassis, node, rack, switch, PDU e site.

Rede

Para HCI, some tráfego de cliente, réplica, recovery e migração. Um cluster com 25 GbE por node pode ter excelente performance normal e ainda sofrer durante backfill + live migration + backup. Se essas operações coexistem, QoS, redes separadas ou 100 GbE podem ser economicamente justificáveis.

CAPÍTULO 18

Arquiteturas de referência

Três desenhos para transformar requisitos em componentes concretos.

Cenário A - 3 nodes HCI para produção

Camada	Recomendação de referência
Compute	3 servidores, 1-2 sockets, RAM ECC, fontes redundantes
Boot	2 SSDs em mirror
Ceph	4-8 SSD/NVMe enterprise por node, JBOD/HBA IT mode
Network	2×25 GbE ou 2×100 GbE para dados; rede de gestão separada
Corosync	links redundantes; não competir com recovery Ceph
Backup	PBS externo ao cluster + cópia off-site
HA	Guests críticos registrados e capacidade N-1 validada

Cenário B - 5 nodes com SAN enterprise

Use PVE como compute cluster e mantenha storage compartilhado em array HA Fibre Channel/iSCSI. Isso reduz complexidade Ceph e pode ser ideal quando a empresa já domina SAN, tem investimentos existentes e exige features específicas do array.

Cenário C - 2 nodes + qdevice no edge

Dois nodes com qdevice resolvem quorum, mas não criam capacidade mágica. Use storage compartilhado redundante ou ZFS replication com RPO explícito. Esse desenho é útil em filiais/edge quando três servidores completos são economicamente inviáveis.

Multi-site

Mantenha clusters locais por site para latência e quorum. Use PDM para visão e operações globais, PBS Sync para cópias off-site e, quando aplicável, cross-cluster migration. Não trate dois sites distantes como um único cluster apenas para “ter HA geográfico”.

CAPÍTULO 19

Migração de VMware/Hyper-V e modernização

Migração não é apenas converter disco: envolve rede, drivers, CPU, backup, observabilidade e rollback.

Do VMware

O PVE possui mecanismos de importação de guests ESXi/OVA em versões recentes. Mesmo assim, a estratégia de produção deve mapear vNICs, VLANs, storage, snapshots, VMware Tools → QEMU Guest Agent/VirtIO, BIOS/UEFI e tipos de controladora. Teste cada classe de workload antes da onda de migração.

Do Hyper-V

VMs Hyper-V normalmente exigem conversão/importação de discos VHD/VHDX e ajuste de drivers/controladoras. Para Windows, prepare VirtIO e valide boot, rede, activation e aplicações. Não faça a primeira validação em uma janela de produção.

Fases recomendadas

1. Discovery: inventário, dependências, RTO/RPO, VLANs, storage e licenças.
2. Landing zone: cluster PVE, redes, storage, DNS/NTP, identidade, backup e observabilidade.
3. Pilot: workloads não críticos e testes de performance/rollback.
4. Waves: migração por grupos de dependência, com janela e critérios de sucesso.
5. Optimization: CPU models, VirtIO, cache, backup, HA e rightsizing.
6. Decommission: somente após retenção/rollback e validação de negócio.

CAPÍTULO 20

Runbook operacional, testes de falha e checklist

Produção madura transforma operações raras em procedimentos ensaiados.

Teste de aceite do cluster

Teste	Resultado esperado
Desligar 1 link Corosync	Cluster mantém membership pelo link redundante
Desligar 1 node	Quorum permanece; guests HA recuperam conforme política
Falhar 1 OSD	Ceph permanece disponível e inicia recovery
Falhar 1 switch de dados	Bond/rede redundante mantém tráfego
Live migrate VM crítica	Sem perda de sessão além do comportamento esperado da aplicação
Restore PBS	VM restaurada e aplicação validada dentro do RTO
IdP indisponível	Conta local break-glass consegue administrar o ambiente

Patching

1. Validar backup e saúde de cluster/Ceph.
2. Migrar/evacuar workloads do node.
3. Entrar em maintenance workflow quando aplicável.
4. Aplicar updates e reboot se necessário.
5. Validar NICs, Corosync, storage, HA e versões.
6. Reintroduzir node e observar recovery/rebalance.
7. Repetir node a node, nunca “tudo de uma vez” sem desenho específico.

Mudanças de configuração

- Snapshot não é rollback universal: mudanças de rede/cluster/storage podem estar fora da VM.
- Tenha change ticket, diff de configuração e plano de retorno.
- Automatize repetição, mas mantenha gates humanos para mudanças de alto risco.
- Guarde export/backup de configuração e documentação de switch/SAN/BMC.

CAPÍTULO 21

Erros comuns e decisões de projeto

Os incidentes mais caros costumam começar com uma premissa simplificada demais.

Erro	Por que falha	Correção
“Cluster = HA”	Sem storage e HA config, guest não recupera sozinho	Configurar HA, quorum e storage/repl.
2 nodes sem qdevice	Perda de um voto elimina maioria	Adicionar qdevice ou terceiro node
Ceph em 1/10 GbE com NVMe	Rede vira gargalo e recovery compete com produção	25/100 GbE e teste degradado
Ceph = backup	Replica exclusão/ransomware/corrupção	PBS externo + off-site/offline
80-90% de RAM normal	N-1 não cabe após falha	Capacidade reservada e rightsizing
SSO sem break-glass	Falha do IdP bloqueia administração	Conta local forte e testada
IaC + mudanças manuais	Drift e resultados não determinísticos	Ownership e pipeline definidos
Backup sem restore	Sucesso de job não garante recuperabilidade	Restore drills e RTO medido

Pergunta que deve anteceder a tecnologia
Qual falha o negócio exige sobreviver - disco, node, switch, rack, site, operador, ransomware - e em quanto tempo? A resposta define o desenho mais do que o nome do hypervisor.

CAPÍTULO 22

Conclusão, glossário e referências

Proxmox é simples de começar e suficientemente profundo para exigir engenharia disciplinada em produção.

Um ambiente Proxmox robusto não é “três servidores com uma interface bonita”. É um sistema distribuído com dependências explícitas: quorum, rede, storage, identidade, backup, capacidade de failover, observabilidade e processos. O ganho é flexibilidade: a mesma plataforma pode operar com ZFS local, SAN tradicional ou Ceph HCI, em x86-64 ou Arm64 suportado, administrada por GUI, API ou IaC.

Resumo de decisões

- 3+ nodes para produção HA é a referência mais simples; 2 nodes exige qdevice e desenho cuidadoso.
- Ceph é excelente quando há escala, rede rápida, discos adequados e capacidade operacional.
- Storage compartilhado tradicional pode ser a melhor solução quando simplicidade e SAN madura pesam mais.
- PDM adiciona gestão multi-site sem tornar-se dependência para execução das VMs.
- Entra ID via OIDC moderniza SSO, mas RBAC/ACL e break-glass continuam indispensáveis.
- Subscriptions são por socket e não bloqueiam features; escolha o nível pelo suporte/SLA e política corporativa.
- Benchmarks devem virar testes locais de aceite, inclusive em estado degradado.

Glossário essencial

Termo	Definição curta
PVE	Proxmox Virtual Environment
PBS	Proxmox Backup Server
PDM	Proxmox Datacenter Manager
KVM	Virtualização de hardware no kernel Linux
LXC	Containers de sistema Linux

Termo	Definição curta
pmxcfs	Filesystem de configuração distribuída do cluster
Corosync	Membership/mensageria/quorum do cluster
QDevice	Voto externo para auxiliar quorum
OSD	Daemon de storage do Ceph
RBD	Bloco distribuído Ceph usado para discos de VM
CRUSH	Algoritmo de placement do Ceph
RPO	Máxima perda de dados tolerável no tempo
RTO	Tempo alvo para restaurar o serviço
OIDC	OpenID Connect, protocolo de identidade sobre OAuth 2.0
RBAC	Controle de acesso baseado em papéis

Fontes principais

Data de corte desta edição: 02 de setembro de 2026. Preços, SLA e versões devem ser revalidados antes de contratação ou projeto.

Fonte	Referência
Apresentação-base do usuário	Cluster Proxmox: Do Zero ao Avançado, 18 slides.
Proxmox VE Overview / Features	https://www.proxmox.com/en/products/proxmox-virtual-environment/overview | /features
Proxmox VE 9.2 release	https://www.proxmox.com/en/about/company-details/press-releases/proxmox-virtual-environment-9-2
Proxmox VE pricing	https://www.proxmox.com/en/products/proxmox-virtual-environment/pricing
Proxmox VE documentation 9.x	https://www.proxmox.com/en/downloads/proxmox-virtual-environment/documentation
User Management / OpenID Connect	https://github.com/proxmox/pve-docs/blob/master/pveum.adoc
Ceph Benchmark 2023/12	https://www.proxmox.com/en/downloads/proxmox-virtual-environment/documentation/proxmox-ve-ceph-benchmark-2023-12
Proxmox Backup Server 4.2 docs	https://pbs.proxmox.com/docs/
Proxmox Datacenter Manager 1.1	https://pdm.proxmox.com/docs/ | https://www.proxmox.com/en/about/company-details/press-releases/proxmox-datacenter-manager-1-1
Arm64 support	https://www.proxmox.com/en/about/company-details/press-releases/proxmox-virtual-environment-launches-official-arm64-support
24/7 support announcement	https://www.proxmox.com/en/about/company-details/press-releases/proxmox-24-7-support-and-proxmox-north-america