AYI NEDJIMI Consultants

Architecture Proxmox VE 9.1 : Cluster 3 Nœuds + HA

Catégorie : Virtualisation Lecture : 5 min Publié le : 22/03/2026 Auteur : Ayi NEDJIMI

Architecture complète cluster Proxmox VE 9.1 en 3 nœuds : MLAG/LACP, Corosync Kronosnet, Ceph NVMe, PVE Firewall 3 niveaux et meilleures pratiques.

La conception d'un **cluster Proxmox VE 9.1** en **3 nœuds** requiert une approche méthodique intégrant les meilleures pratiques en matière de réseau, stockage et sécurité. Ce guide architecturel complet couvre le dimensionnement **MLAG/LACP** pour la redondance réseau, la configuration *Corosync (moteur de communication du cluster Proxmox)* avec le protocole **Kronosnet** pour la résilience du quorum, le stockage distribué **Ceph NVMe** pour des performances I/O optimales, et la mise en place d'un pare-feu **PVE Firewall** en 3 niveaux (datacenter, nœud, VM) pour une infrastructure de production robuste. Ce guide s'adresse aux administrateurs souhaitant déployer une infrastructure haute disponibilité avec un budget maîtrisé, en exploitant pleinement les capacités natives de Proxmox VE 9.1 sans dépendance à des solutions propriétaires. Chaque section apporte des exemples de configuration concrets, des ratios de dimensionnement et des checklists de validation.

Points clés à retenir

- Un cluster 3 nœuds est le minimum recommandé pour la haute disponibilité : il garantit le quorum même avec la perte d'un nœud (2/3 votes disponibles).
- Séparer les réseaux Corosync, migration et stockage Ceph sur des interfaces dédiées est impératif pour éviter la congestion et les pannes de quorum.
- Le stockage Ceph NVMe en hyper-convergence offre des performances optimales, mais requiert au minimum 3 OSDs par nœud et des réseaux publics/cluster séparés.
- Le PVE Firewall en 3 niveaux permet une politique de sécurité granulaire : règles datacenter globales, règles par nœud et règles par VM/CT.

Architecture Réseau du Cluster : MLAG, LACP et Séparation des Flux

L'architecture réseau d'un cluster **Proxmox VE 9.1** en production nécessite une séparation stricte des flux pour éviter les interférences entre le trafic de gestion, **Corosync**, la **live migration** et le stockage **Ceph**. La configuration recommandée utilise au minimum 4 interfaces physiques par nœud :

- **vmbr0** (Management + VM) : 1GbE ou 10GbE, bonding LACP actif/passif
- **vmbr1** (Corosync + Migration) : 10GbE dédié, faible latence (< 2ms)
- **vmbr2** (Ceph Public Network) : 10/25GbE pour les accès RBD clients

• **vmbr3** (Ceph Cluster Network) : 10/25GbE pour la réplication interne Ceph

Le *MLAG (Multi-Chassis Link Aggregation)* avec deux switches ToR permet d'éliminer les SPoF (Single Points of Failure) au niveau des commutateurs. Le **LACP (802.3ad)** agrège deux liaisons physiques en un canal logique, doublant la bande passante et assurant la redondance. La configuration dans **/etc/network/interfaces** utilise le module **bonding** Linux avec `bond-mode 802.3ad`.

Corosync et Kronosnet : Gestion du Quorum et Résilience

Corosync est le moteur de communication du cluster Proxmox, responsable de la gestion du *quorum (mécanisme de vote majoritaire empêchant le split-brain)*. Dans Proxmox VE 9.1, Corosync utilise **Kronosnet** comme couche de transport, offrant chiffrement des communications (AES-256), compression et support multi-liens.

Pour un cluster 3 nœuds, le quorum est atteint avec 2 nœuds actifs (quorum = N/2 + 1 = 2). La configuration recommande deux anneaux Corosync sur des interfaces réseau distinctes pour la redondance. Le fichier **/etc/pve/corosync.conf** définit les membres du cluster, les adresses de chaque anneau et le paramètre **expected_votes**.

Commandes de diagnostic essentielles : **pvecm status** (état du cluster et quorum), **corosync-quorumtool -s** (détail du vote), **journalctl -u corosync -f** (logs en temps réel). Pour la gestion avancée du cluster, consultez notre guide d'administration CLI Proxmox.

Stockage Ceph NVMe en Hyper-Convergence

Ceph est la solution de stockage distribué native de Proxmox VE, offrant un stockage bloc (**RBD**), objet et fichier hautement disponible. En configuration hyper-convergée avec des disques **NVMe**, Ceph atteint des performances I/O exceptionnelles : latence < 1ms, IOPS > 500k par OSD.

L'architecture recommandée pour 3 nœuds : 3 OSDs NVMe par nœud (total 9 OSDs), factor de réplication **size=3, min_size=2**, et 2 MONs/MGRs sur les nœuds principaux. Les *Placement Groups (PGs)* doivent être calculés selon la formule : PGs = (OSDs × 100) / facteur_réplication. Pour 9 OSDs avec réplication 3 : PGs = 9 × 100 / 3 = 300 PGs.

La séparation des réseaux Ceph Public (accès clients RBD) et Cluster (réplication interne) est obligatoire en production. Le wiki Proxmox Ceph détaille les étapes de déploiement pour éviter que le trafic de réplication ne sature la bande passante des VMs. Pour le dimensionnement complet, consultez notre guide de dimensionnement Proxmox VE 9.

PVE Firewall : Sécurité en 3 Niveaux

Le **PVE Firewall** s'applique à 3 niveaux hiérarchiques : **Datacenter** (règles globales pour tout le cluster), **Nœud** (règles spécifiques à chaque hôte Proxmox) et **VM/CT** (règles par machine virtuelle ou conteneur). Cette architecture permet d'appliquer des politiques de sécurité granulaires sans duplication de règles.

Configuration de base recommandée : politique par défaut DROP sur le datacenter, avec règles d'autorisation explicites pour SSH (port 22), Web UI (port 8006), Corosync (UDP 5405-5406) et migrations (ports 60000-60050). Les **IPSets** regroupent les plages d'adresses de confiance (réseaux admin, supervision). Pour une sécurisation complète, référez-vous à notre guide de hardening Proxmox VE.

Composant	Spécification minimale	Recommandée production
CPU par nœud	8 cœurs / 16 threads	32 cœurs / 64 threads
RAM par nœud	64 Go ECC	256 Go ECC
Stockage OS	2× SSD SATA RAID1	2× NVMe ZFS mirror
Réseau Corosync	1GbE dédié	10GbE dédié redondant
Réseau Ceph	10GbE	25GbE public + cluster

Haute Disponibilité : HA Manager et Fencing

Le **PVE HA Manager** surveille l'état des VMs et CTs protégés par HA et déclenche automatiquement le failover en cas de panne d'un nœud. Le *fencing (STONITH - Shoot The Other Node In The Head)* est le mécanisme critique qui éteint physiquement un nœud défaillant avant de redémarrer ses VMs ailleurs, éliminant le risque de split-brain avec le stockage partagé.

La configuration du fencing dans Proxmox utilise des dispositifs IPMI (iDRAC, iLO, IPMI 2.0) ou des PDU réseau (APC, Raritan). Sans fencing configuré, le HA Manager attend un timeout avant de redémarrer les VMs, augmentant le temps de récupération. Les groupes HA permettent de définir des priorités de nœuds pour le placement des VMs après failover.

Pour approfondir la réplication des données entre nœuds, consultez notre guide de réplication ZFS Proxmox. Pour automatiser le déploiement de votre cluster, voir notre guide de déploiement automatisé Proxmox.

Questions fréquentes

Pourquoi un cluster Proxmox doit-il avoir un nombre impair de nœuds ?

Le **quorum Corosync** repose sur un vote majoritaire : pour qu'une décision soit prise (démarrer/arrêter des VMs, modifier la configuration), plus de la moitié des nœuds doivent être disponibles. Avec 3 nœuds, la perte d'un nœud laisse 2/3 des votes disponibles (quorum atteint). Avec 2 nœuds, la perte d'un seul nœud bloque le cluster car aucun nœud n'a la majorité (1/2 insuffisant). Un nombre impair garantit toujours une majorité possible, évitant les situations de split-brain où deux partitions agissent indépendamment.

Comment séparer efficacement les réseaux Corosync et stockage dans Proxmox VE 9.1 ?

La séparation des réseaux se configure dans **/etc/network/interfaces** en créant des bridges dédiés pour chaque flux. Corosync utilise les adresses définies dans **/etc/pve/corosync.conf** (paramètre `ring0_addr` et `ring1_addr`). Le stockage Ceph utilise les paramètres `public_network` et `cluster_network` dans **/etc/ceph/ceph.conf**. Une bonne pratique consiste à utiliser des VLANs dédiés sur une infrastructure switch redondante, avec des interfaces physiques distinctes ou des bonds LACP pour chaque type de trafic.

Quelle est la configuration minimale pour un cluster Proxmox VE 9.1 en production ?

Pour une infrastructure de production fiable, chaque nœud doit disposer d'au minimum : **32 Go RAM ECC** (ECC obligatoire pour ZFS), **2 interfaces réseau 10GbE** (gestion + Corosync/migration), **1 interface 10GbE** pour Ceph, et **2 SSD NVMe** en mirror ZFS pour l'OS. Au niveau cluster, au moins **3 nœuds** pour le quorum, un **device de fencing IPMI** sur chaque nœud et une **alimentation redondante** sur les serveurs. La documentation officielle Proxmox VE détaille les prérequis matériels selon les versions.

Sources et références : Proxmox VE Wiki · ANSSI

Conclusion

Une architecture **Proxmox VE 9.1** en cluster 3 nœuds bien conçue offre haute disponibilité, performances et sécurité sans dépendance à des solutions propriétaires. La séparation des réseaux, le stockage Ceph hyper-convergé et le PVE Firewall multicouche constituent les piliers d'une infrastructure production-ready. L'investissement dans une architecture solide dès le départ évite les refontes coûteuses et les incidents en production.

Ayi NEDJIMI Consultants — Expert cybersécurité offensive & intelligence artificielle

ayinedjimi-consultants.fr · ayi@ayinedjimi-consultants.fr