WHOLESALE WINERY Tour

# Scalare Proxmox VE

**Michele Gallo**
**Head of Datacenter Operations**

# Proxmox VE at Scale

WHOLESALE WINERY Tour

## Dal Cluster singolo alla piattaforma Multi DataCenter

Architettura, Performance & Casi d'Uso Reali

**Technical Director, Proxmox Cloud**
Proxmox Clustering: ★★★★★★★★★☆ (9/10)

Network Architecture: ★★★★★★★★☆☆ (8/10)

Ceph Storage: ★★★★★★★★☆☆ (8/10)


Michele Gallo, Head of Datacenter Operations, RaiWay

CONTESTO 2026

# Perchè questa conversazione è importante

Il panorama delle infrastrutture aziendali è cambiato radicalmente

**3–5×**

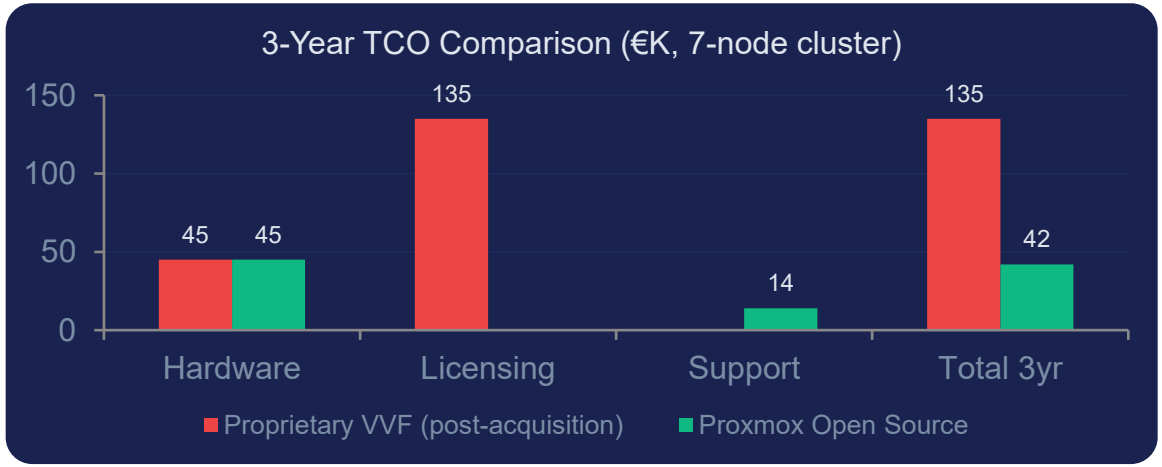
Aumento dei costi di licenza dopo l'acquisizione di Broadcom

**€100K+**

Barriera di investimento per le PMI con stack proprietari

**8 Anni**

Maturità di Proxmox 2018 → 2026




⚠ Vincolo del fornitore: API, formati e strumenti proprietari

✓ Maturità dell'open source: Proxmox 9.x è di livello enterprise

✓ Comunità attiva + opzioni di supporto commerciale disponibili

→ Dual socket – 32C – 384G ram – 6x 1,92TB NVMe

VVF: restrizioni vSAN/SDN, sconto del 10% sul prezzo di listino
Proxmox Premium: riduzione del 50% rispetto al prezzo STANDARD

ARCHITETTURA

# Dimensionamento dei cluster: la matematica del quorum

**Quorum = ⌊N/2⌋ + 1**

Numero minimo di voti necessari per il funzionamento del cluster

| Nodes (N) | Quorum | Guasti tollerati | Use Case | Livello di rischio |
|---|---|---|---|---|
| 3 | 2 | 1 | Dev/small prod | ⚠ Medio |
| 5 | 3 | 2 | PMI produzione | ✓ Buono |
| 7 | 4 | 3 | MSP / cloud provider | ✓✓ Raccomandato |


💡 Utilizzare sempre un numero dispari di nodi per evitare lo split-brain. Non eseguire mai un cluster a 2 nodi in produzione senza un QDevice.

SCALING

# Strategie di scalabilità: quando è opportuno crescere orizzontalmente

WHOLESALE WINERY Tour

Tre modelli architetturali: scegli in base al profilo del carico di lavoro e al numero di nodi

## HCI (Hyperconverged)

*≤11 nodi*


- Compute + storage su ogni nodo
- Ceph si distribuisce su tutti gli host
- Cablaggio più semplice, maggiore densità
- PRO: simile al cloud, scalabile orizzontalmente

## Converged

*11–20 nodi*


- Mix: alcuni nodi storage+compute
- Nodi con elevata capacità di archiviazione vs nodi con elevata capacità di calcolo
- Ceph è ancora distribuito, ma a livelli
- PRO: workload misto tipo MSPs- (VPS business)

## Disaggregated

*20+ nodi*


- Nodi solo storage (Ceph OSDs)
- Nodi solo compute (VMs only)
- Massima ottimizzazione per ruolo
- PRO: team specializzati e di grandi dimensioni

Evoluzione: 7-node HCI → 11-node converged (7 HCI + 4 compute-only) → 20-node disaggregated (8 storage + 12 compute)

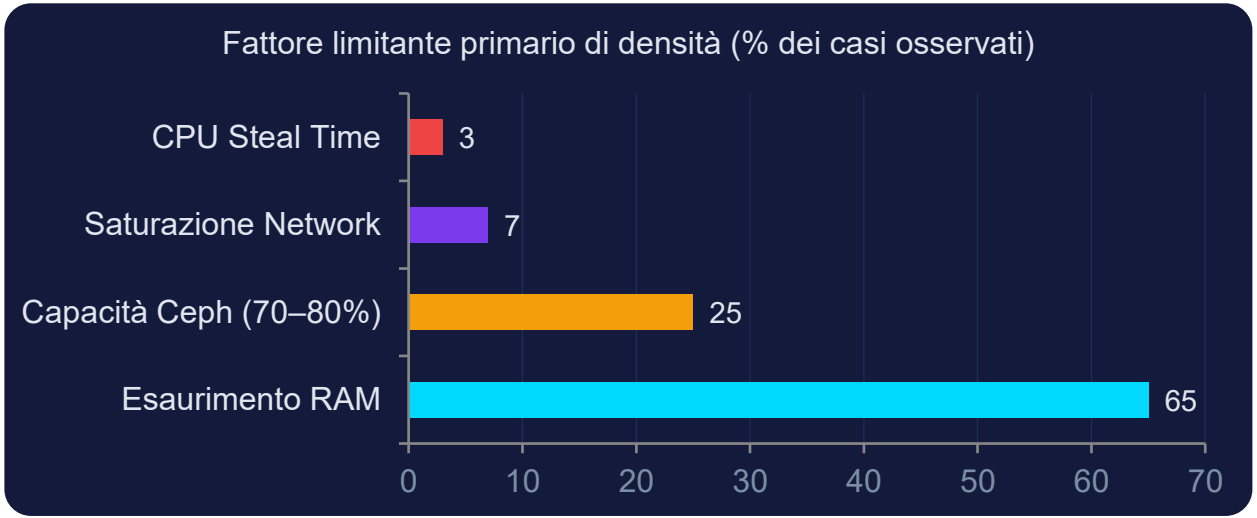
DENSITA’

# Cosa limita la densità delle macchine virtuali? (Spoiler: non la CPU)

Vincoli di densità reali nei cluster di produzione multi-tenant

**25**

VMs / host
Media in Produzione




**20–30 VMs**

Typical workloads (2–8 vCPU, 4–16GB RAM)

**8–15 VMs**

Memory-heavy (DBs, analytics)

💡 L'overcommit della CPU con un rapporto 4:1 è sicuro. L'overcommit della memoria NON lo è: i driver balloon aiutano, ma causano picchi di latenza.

WHOLESALE WINERY Tour

# Analisi approfondita di Ceph: Tiering e prestazioni

## CEPH key components

→ Minimum 4GB ram, 6GB for NVMe OSD preferred
→ Approximately 0.5 to 1 CPU core per OSD under load
→ Network encryption (if enabled) adds 10-15% CPU overhead
→ MON : Maintains cluster map, requires quorum (typically 3 or 5)
→ MON + MGR : consider 10% overhead for CPU allocation

## NVMe Pool — Tier 1

→ Database VMs, Redis, real-time analytics
→ Low-latency SLA: <2ms p99
→ CRUSH rule: NVMe failure domain only

## SSD Pool — Tier 2

→ General-purpose VMs, web, app servers
→ Balanced IOPS: 15–30K
→ CRUSH rule: host-level failure domain

**Latency Under Load (fio, 70% read/write mix)**

| Percentile | NVMe OSDs | SSD OSDs |
|---|---|---|
| p50 (median) | 0.45 ms | 1.9 ms |
| p95 | 1.2 ms | 8.1 ms |
| p99 | 2.9 ms | 21.4 ms |
| p99.9 | 8.5 ms | 45+ ms |

⚠ La latenza p99+ è il punto debole dei cluster SSD: i carichi di lavoro dei database ne risentono immediatamente.

✓ Ceph WAL/DB: Mantieni sempre NVMe anche con OSD SSD

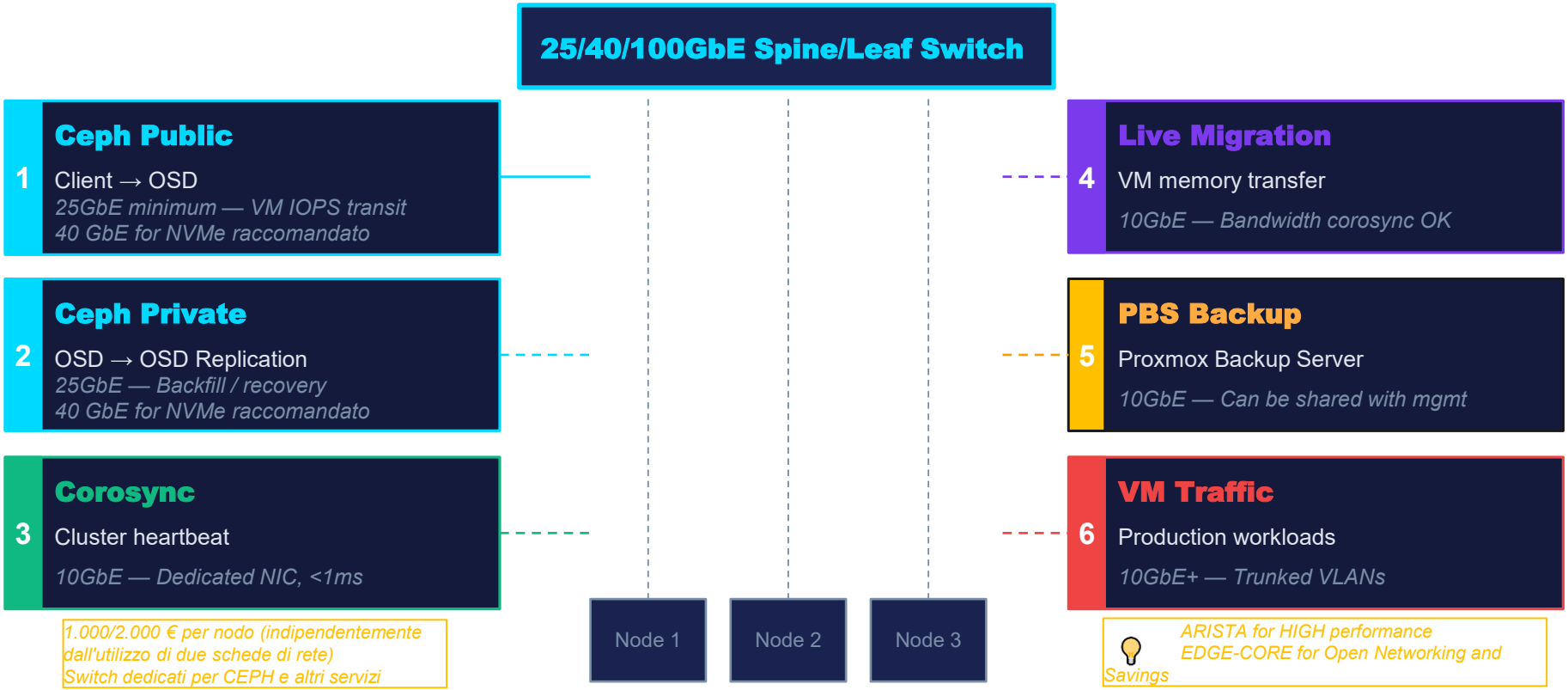
**Replica x3:** Massime prestazioni, ripristino semplice, costo di archiviazione triplo. **Erasure Coding (8+3):** Efficienza di capacità del 62%, maggiore overhead della CPU, latenza circa doppia: ideale per archiviazione a freddo o pool di backup.

NETWORKING

# Architettura di Rete: Il Setup Perfetto

WHOLESALE WINERY Tour

Sei piani di traffico separati: requisiti di larghezza di banda e motivazioni



CASO REALE

# Compromessi ottimizzati in termini di costi.

Quando i vincoli di bilancio incontrano i requisiti di produzione: cosa funziona davvero?

## ✓ Setup Perfetto (6 networks)

€1,500 per nodo in NICs + switches (CEPH only 2x €8000)

- 25GbE × 2 → Ceph Public + Private (separate)
- 10GbE × 1 → Corosync (dedicated, no VLAN)
- 10GbE × 1 → Live Migration (dedicated)
- 10GbE × 1 → PBS Backup + Management
- 10GbE+ × 1 → VM Production Traffic (trunked)

## ⚡ Budget Setup (3 networks)

€400 per nodo — risparmi ~€1.1K/nodo, €23K on 7-node cluster

### 25GbE bonded (LACP)

Ceph Public + Private shared
⚠ Recovery storms possono saturarla

### 10GbE (VLAN 10)

Corosync + Live Migration
⚠ La migrazione rallenta durante gli eventi di cluster

### 10GbE (VLAN 20/30)

VM Traffic + PBS
Adatto alla maggior parte dei carichi di produzione

**Budget setup funziona quando:** carico di archiviazione sostenuto <50% | migrazioni live simultanee <10 | nessun ripristino di Ceph durante l'orario lavorativo

MULTI-DC

# BGP/EVPN: Cluster Multi-Datacenter

Migrazione live di macchine virtuali tra data center: requisiti architetturali

## Use Case #1 - BC

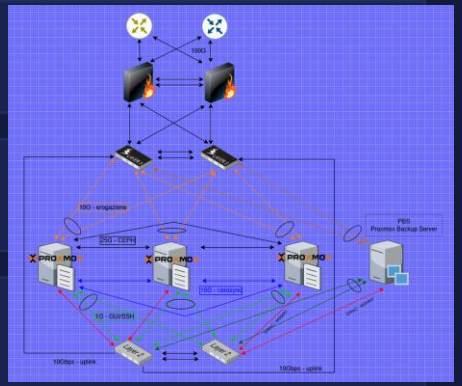
Business continuity with active-active sites
Zero-downtime datacenter maintenance

## Use Case #2 - DR

Two DCs 50 km apart, 2ms latency via dark fiber

### Datacenter A (Primary)

nodo-01
nodo-02
nodo-03



*BGP/EVPN (AS 65001)*

Streched VLAN

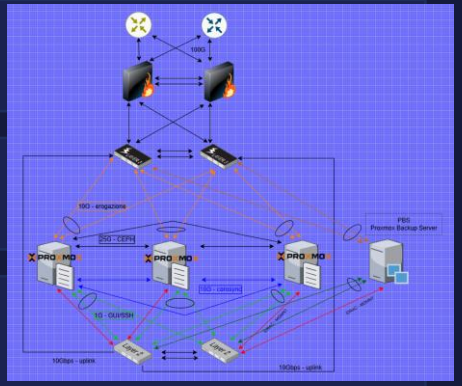
BGP
Core
Routing

**AS 65000**

Storage replica

### Datacenter B (DR)

nodo-04
nodo-05
nodo-06



*BGP/EVPN (AS 65002)*

## <5ms RTT

Requisiti di latenza tra data center per la migrazione live
Larghezza di banda: minimo 10 GbE, preferibilmente 40 GbE

## Anycast IPs

Indirizzi IP pubblici annunciati simultaneamente da entrambi i data center.

## Ceph Stretch

È necessario un terzo sito arbitro per evitare lo split-brain nei cluster distribuiti

# High Availability: cosa succede realmente?

Cronologia dei guasti dei nodi: percorsi di ripristino con HA abilitato rispetto a percorsi di ripristino manuali

## HA ABILITATA

- **T+0s** — Node kernel panic / hardware failure
- **T+2s** — Corosync rileva assenza di heartbeat
- **T+5s** — Quorum ricalcolato, voti dei nodi ritirati.
- **T+10s** — Fencing: IPMI power-off or watchdog fires
- **T+30s** — Il gestore HA sposta le macchine virtuali sui nodi rimanenti
- **T+45s** — Avvio di macchine virtuali su nuovi nodi

## ⚠ HA NON ABILITATA— Ripristino manuale

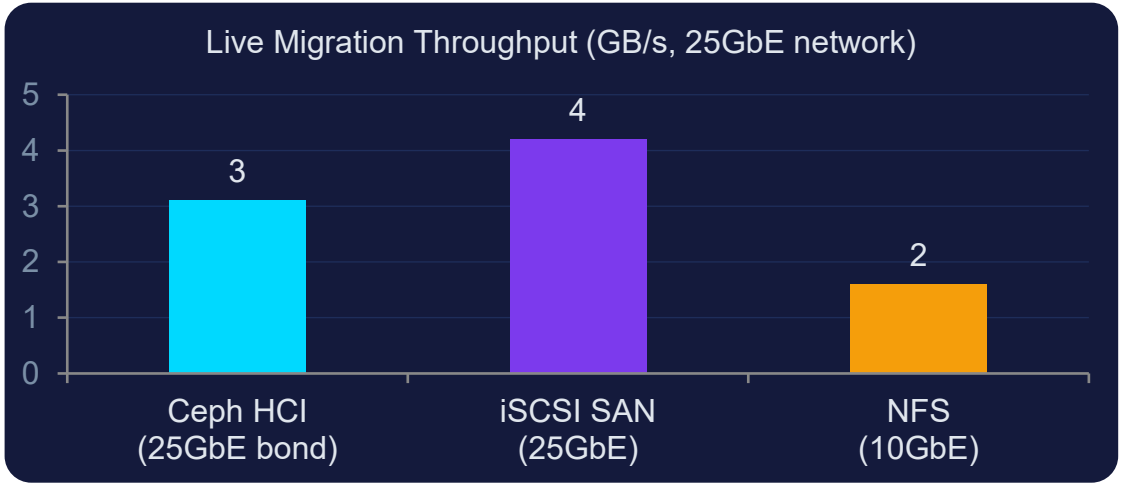
1. Rilevare i guasti (avviso di monitoraggio o reclamo dell'utente)
2. Connettiti tramite SSH al nodo funzionante
3. Arrestare forzatamente le macchine virtuali ancora contrassegnate come in esecuzione
4. Identificare quali VM si trovavano sul nodo guasto
5. Avviare manualmente le macchine virtuali sui nodi rimanenti
6. Aggiornare il DNS/il bilanciatore di carico, se necessario.
7. **Total downtime: 5–30 minuti è tipico**

💡 Abilita sempre l'alta disponibilità (HA) per le macchine virtuali di produzione. Il tempo di ripristino (RTO) di 30 secondi è gestibile. Il ripristino manuale di 30 minuti, invece, non lo è.

MIGRATION

# Live Migration Performance

Velocità di migrazione in base al backend di archiviazione e finestre di inattività effettive delle macchine virtuali




***La larghezza di banda della rete è il collo di bottiglia, non lo spazio di archiviazione.***

**VM Downtime During Live Migration**

| VM Tipo | RAM | Dirty Rate | Down time |
|---|---|---|---|
| Small web server | 4 GB | Low (<50 MB/s) | < 1 sec |
| App server | 32 GB | Medium (~200 MB/s) | 1–2 sec |
| Database server | 128 GB | High (>500 MB/s) | 3–6 sec |
| Memory-intensive cache | 256 GB | Very high (>1 GB/s) | Stop & move |

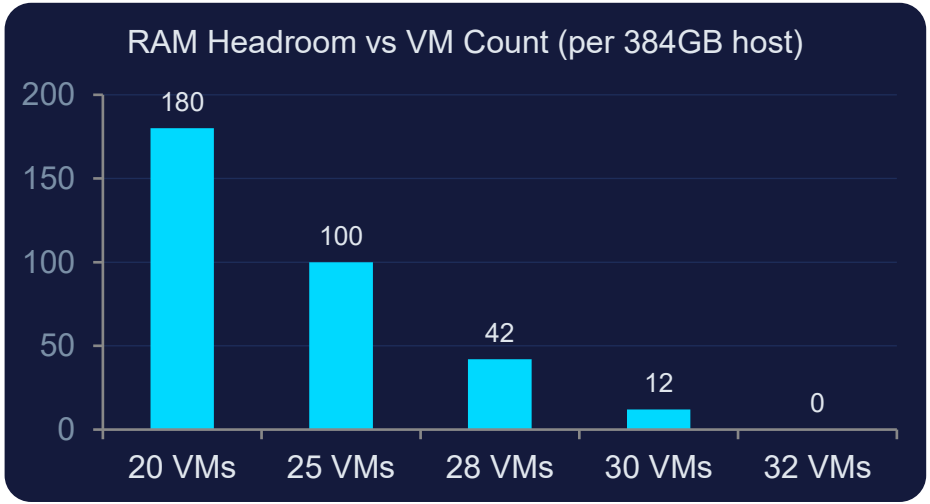
💡 High dirty-page VMs (busy DBs): pianifica le migrazioni durante le finestre di basso carico o utilizza la migrazione offline per istanze da 256 GB o superiori..

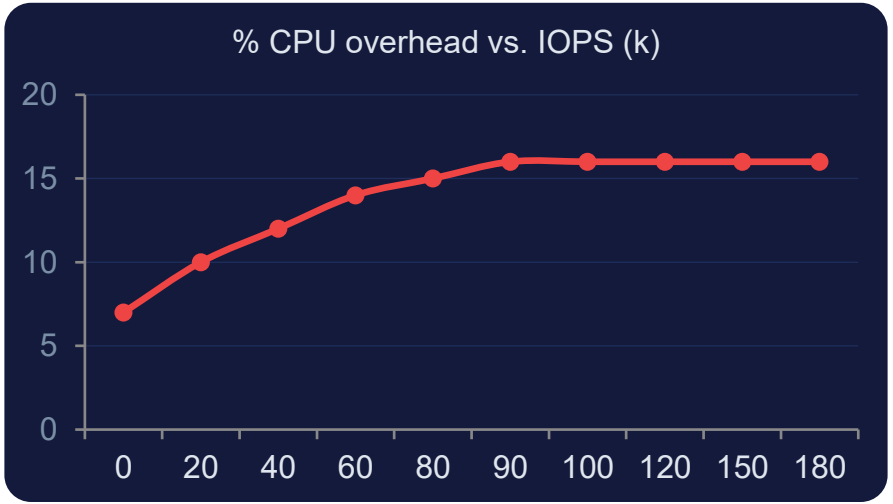
✓ Vantaggio di Ceph HCI: nessun trasferimento di dati di archiviazione durante la migrazione, viene trasferita solo la RAM.

TUNING

# Colli di bottiglia nelle prestazioni: cosa ho riscontrato

Modelli di esaurimento delle risorse e limitazione del recupero Ceph




→ *RAM limite tipico a ~30 VMs*

→ *Allocare sempre almeno il 15% per CPU overhead*

**CEPH Capacity Behaviour**

**0-60%** Prestazioni ottimali

**60-70%** Impatto Minimo, rebalancing aumenta

**70-85%** Grave degrado, aumento della latenza

**85-95%** CRITICAL, performance calano del 50%+

ECOSYSTEM

# Proxmox vs OpenStack vs Kubernetes

## Proxmox VE

*VM-Centric Virtualization*

**Caso d'uso principale**
Virtual machines, legacy app lift & shift

**Team size**
1–5 ops engineers

**Complessità**
★★☆☆☆ BASSA — GUI-driven

**API maturity**
REST API, Terraform provider

**Scaling model**
Cluster orizzontale + verticale

**Curva d'apprendimento**
Giorni/settimane

**Best for**
SME, MSP, VMware migration

## OpenStack

*IaaS Cloud Platform*

**Caso d'uso principale**
API-driven cloud, multi-tenant IaaS

**Team size**
5–30+ cloud engineers

**Complessità**
★★★★★ ALTA — many components

**API maturity**
Full AWS-like API surface

**Scaling model**
Disegnato per centinaia di nodi

**Curva d'apprendimento**
Mesi/anni

**Best for**
Telcos, large enterprises, ISPs

## Kubernetes

*Container Orchestration*

**Caso d'uso principale**
Stateless containers, microservices

**Team size**
2–20 DevOps engineers

**Complessità**
★★★★☆ ALTA — ecosystem sprawl

**API maturity**
Massive ecosystem, CNCF projects

**Scaling model**
Pod autoscaling, node pools

**Curva d'apprendimento**
Settimane/mesi

**Best for**
Cloud-native apps, CI/CD workloads

🔑 Nella pratica: Proxmox gestisce le VM. Kubernetes viene eseguito all'interno delle VM. OpenStack è per quando diventerai tu stesso un fornitore di servizi cloud.

MIGRATION

WHOLESALE WINERY Tour

# Esodo da VMware (Post-Broadcom)

## 2023–2026: how the enterprise virtualization market fractured

**Nov 2023** — Completata l'acquisizione di Broadcom. Interrotte le licenze perpetue di VMware..

**Feb 2024** — Modello basato esclusivamente su abbonamento. Molti clienti riscontrano un aumento dei costi da 3 a 5 volte superiore al momento del rinnovo.

**Jun 2024** — vSphere Essentials (livello PMI) non è più disponibile. Non è previsto un punto di ingresso a basso costo.

**2025** — Inizia l'ondata di migrazione di massa. Proxmox, Nutanix e Azure Local registrano una rapida adozione.

**2026** — Proxmox diventa di fatto l'alternativa a VMware per il mercato delle PMI e dei Managed Service Provider (MSP).

**Percorso di Migrazione:** virt-v2v per la conversione OVA/VMDK. Prevedere 2-8 ore per VM per dischi di grandi dimensioni. Pianificare finestre di manutenzione.

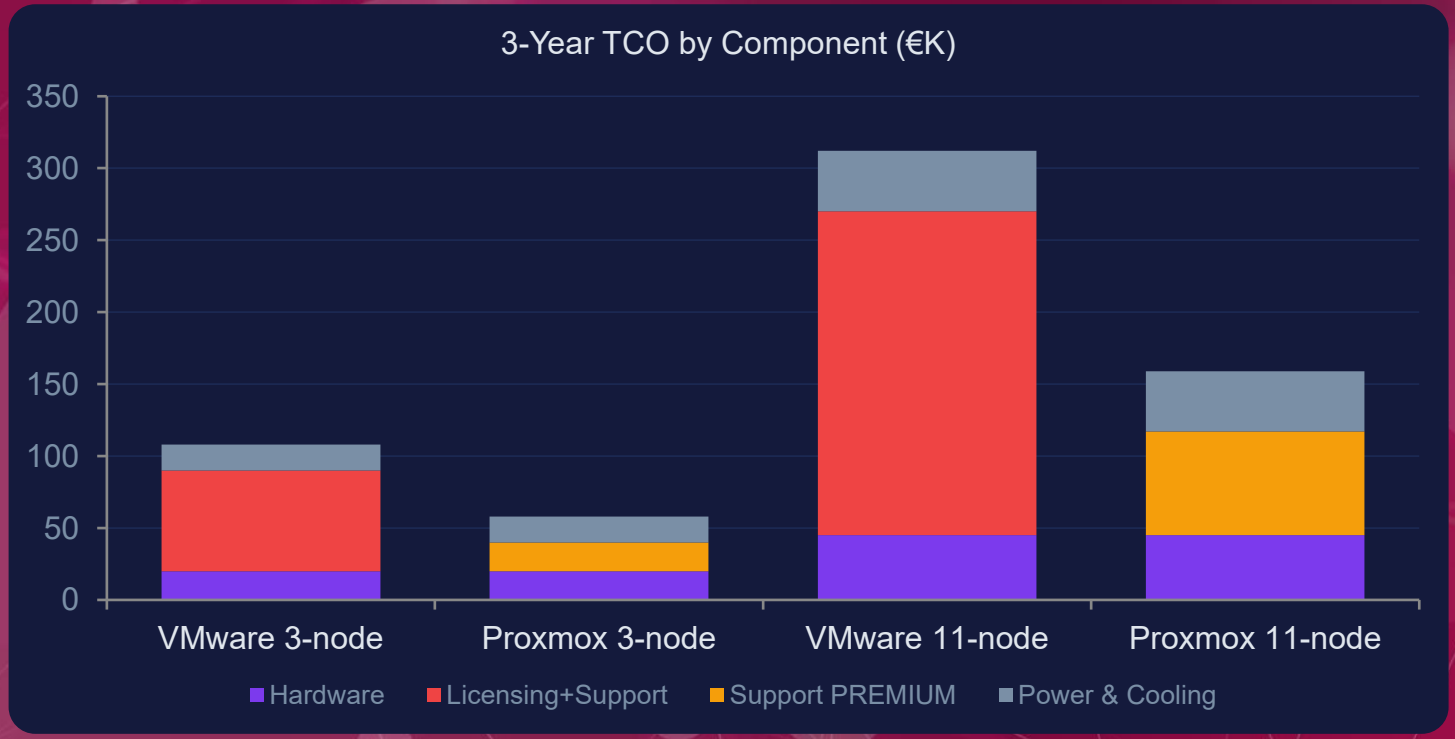
### Confronto funzionalità

| | |
|---|---|
| ✓ | Live migration |
| ✓ | HA failover |
| ✓ | Distributed storage |
| ✓ | VM snapshots |
| ✓ | vGPU support |
| ✗ | vSAN integration (CEPH ✓) |
| ✗ | NSX-T networking (SDN ✓) |
| ✗ | Tanzu (K8s) |

TCO

# TCO: 3-Year Total Cost

Cluster piccolo (3 nodi) vs cluster medio (11 nodi): hardware identico, licenze radicalmente diverse.




**46%**

TCO inferiore

Risparmi ~€50K in 3 anni

**51%**

TCO inferiore

Risparmi ~€156K in 3 anni

* Abbonamento Proxmox Enterprise incluso per il supporto SLA. Costi hardware identici (stesse specifiche del server). Costi energetici identici.

# Domande?

Felice di approfondire su :

01 Dimensionamento del cluster e progettazione del quorum per il tuo specifico carico di lavoro

02 Strategia di migrazione VMware: tempistiche, strumenti e mitigazione dei rischi.

03 Ceph o SAN esterna per il tuo caso d'uso specifico

04 Revisione della progettazione di rete e calcolo della larghezza di banda

05 Architettura di backup: PBS, S3, politiche di conservazione

WHOLESALE WINERY Tour

# THANK YOU!