Migrazione del cluster di macchine virtuali di OAS da OVS a Proxmox VE

Autore: Alessandro Tacchini

Introduzione

Nel nostro istituto, l'OAS-INAF di Bologna, si è reso necessario il passaggio del cluster di virtualizzazione realizzato con Oracle Virtual Server ad un nuovo cluster realizzato con Proxmox Virtual Environment
Da molti anni oramai alcuni servizi informatici, di IASFBO prima ed OAS ora, vengono implementati attraverso server realizzati con macchine virtuali.
I vantaggi sono molteplici: flessibilità, robustezza, velocità nel deployment di una macchina.
Il primo sistema è stato creato utilizzando macchine dismesse dai vari progetti di calcolo e raccolte in un cluster.
Nel tempo, tuttavia, l'aumento dei servizi offerti e l'usura degli apparati hanno portato ad una obsolescenza che ha reso il sistema inutilizzabile.
Da qui la necessità di passare ad un sistema nuovo.
In questo documento si illustrerà il passaggio dal vecchio sistema al nuovo riportando anche l'elenco dei comandi usati.

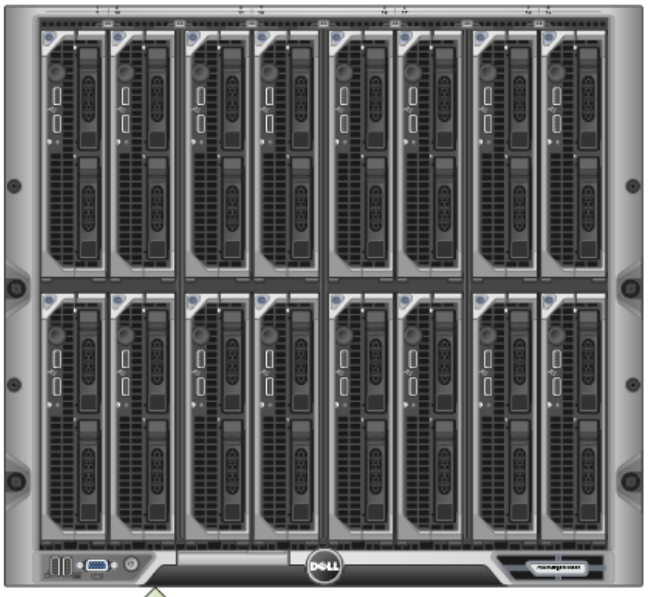
Vecchio Cluster Virtual Machine

È un cluster formato da sette macchine, più una di controllo, non omogenee su cui è installato Oracle VM Server.
Un gruppo di macchine, per la precisione tre, ha 8 Core AMD e 16GB di RAM, due macchine hanno 8 Core Intel e 32 GB di RAM (queste cinque macchine sono contenute all'interno di una Enclosure Dell) ed infine due macchine gemelle hanno 32 Core AMD e 128GB di RAM.

Cluster OVS

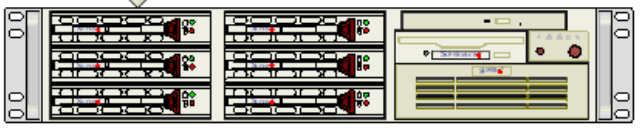
nodo

01 02 03 04 05



Enclosure Dell

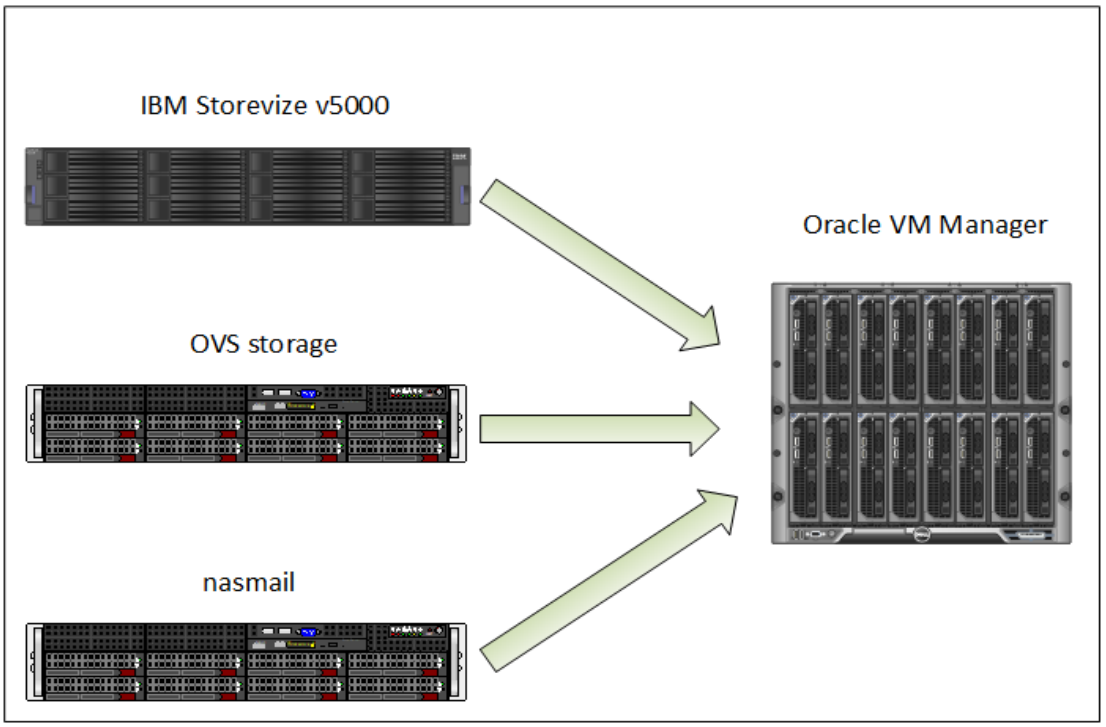
vs1



vs2

Nome nodo	CPU Core	RAM (GB)
nodo01	Intel 8	32
nodo02	Intel 8	32
nodo03	Intel 8	16
nodo05	Intel 8	16
nodo06	Intel 8	16
vs1	AMD 32	128
vs2	AMD 32	128

Lo storage è altrettanto eterogeneo essendo composto da un server Supermicro riciclato da altri progetti di ricerca che esporta via iSCSI circa 6 TB di spazio disco; un server Supermicro che esporta circa 8TB sempre via iSCSI; infine un server SAN IBM Storewize V5000 che esporta via iSCSI 10 TB.
Da notare che OVS gestisce lo storage integrando le fonti esterne ed interne e permette la creazione di repository che vengono poi usati dalle macchine virtuali per lo storage.



Invece in Proxmox VE ogni fonte esterna di storage va agganciata ad ogni nodo del cluster.
Le risorse Hardware quindi sono di dimensioni variabili e questo comporta un sostanziale disequilibrio del cluster che però non incide sulle prestazioni desiderate.
Il software di Oracle realizza un'infrastruttura robusta che offre alta affidabilità ed alta disponibilità.

Le macchine virtuali ospitate sul sistema offrono servizi di due tipi: servizi generali forniti alla struttura e risorse informatiche fornite ai ricercatori nell'ambito, ma non solo, di progetti di ricerca.
I servizi generali erogati sono del tipo: DNS, DHCP, Owncloud, posta elettronica[1], LDAP, RADIUS/EDUROAM, server di licenze[2], gestione delle presenze in caso di emergenze.
L'utilizzo di queste risorse da parte degli utenti può essere di tipo personale, ad esempio utilizzando una macchina virtuale come workstation contenente software scientifico o CAD; oppure può essere legato a progetti di ricerca, alcuni dei quali sono ad esempio: CTA, AGILE[3], THESEUS, EUCLID.
Gli aspetti negativi di OVS (Oracle Virtual Server) sono l'assenza di un sistema di backup integrato e la gestione interna dai file, ad esempio i file che contengono i dischi virtuali sono in formato Raw e sono ubicati in sottodirectory contrassegnate solo da codici numerici. I file stessi sono contrassegnati da codici alfanumerici rendendo difficile l'associazione con la relativa macchina virtuale.
Con una procedura macchinosa è possibile effettuare un disaster recovery avendo a disposizione una copia dei file di immagine dei dischi virtuali e tutte le informazioni relative alle varie macchine virtuali.

[1] Zimbra
[2] Per esempio IDL
[3] Pipeline di Agile

Migrazione ad un cluster Proxmox temporaneo

Poichè il vecchio sistema OVS era diventato talmente instabile da comprometterne l'uso e minarne l'integrità e poiché l'hardware del nuovo cluster Proxmox non era ancora a disposizione si è resa necessaria una azione tempestiva di migrazione ad un cluster temporaneo al fine di mettere al sicuro i dati.
Si è proceduto nel modo seguente: si è scorporata la macchina vs1[4] dal cluster OVS spostando tutte le macchine virtuali e tutti i metadati sui nodi rimanenti, successivamente si è installato Proxmox VE 6.3-3 su un cluster di 4 macchine comprendente la macchina vs1 e tre blade server forniti in prestito da un progetto di ricerca e già presenti nel Centro di Calcolo.
Questi tre nodi sono stati chiamati oasproxmox1, oasproxmox2, oasproxmox3.
Ciascuno di questi nodi ha le seguenti caratteristiche: 16 Core; 2 dischi da 250GB in configurazione zmirror; 24GB di RAM.
Tutti i nodi fanno uso di una rete ethernet a 1GB/s con indirizzi sulla rete nascosta di istituto.
La parte più impegnativa è stata quella di identificare, per ogni macchina virtuale presente su OVS, i file di configurazione ed i codici identificativi dei dischi virtuali, e successivamente di identificare la loro posizione fisica. In pratica è stato trovato il percorso assoluto per ogni file associato ad un disco virtuale.
Questi file sono molto importanti, essi rappresentano l'equivalente del disco fisico di una macchina reale in cui sono contenuti tutti i dati. Costituiscono l'essenza della macchina virtuale.
In OVS tali file hanno l'estensione .img che è del tutto equivalente al comune formato RAW.

[4] evicted

Nome	RAM	CPU/Core	Dischi
LDAP	512MB	1	03deef9a65234aacb3d0f8fc8c76bdf1.img
Lucy	512MB	1	f4846e3cc97a45758e3dc281cfa547aa.img sistema 0004fb0000120000202f3416dd76fa23.img home 0004fb0000120000e9b4efa424a96656.img log
Theseus	4096MB	2	0004fb000012000093d2acb7178ea3c3.img
hangar_web	2048MB	2	db644324a8f34e8a9fb4fc9cf87e1d52.img
hermes	2048MB	2	0004fb00001200004e7c7afe23b68974.img
redsox	4096MB	2	0004fb000012000000135698d14986868.img
Agileref	32246MB	8	0004fb0000120000000524245246ffab36.img sistema
ctasymposium	2048MB	2	0004fb00001200002a09ec084074f457.img
villaserver	8196MB	4	80b6898fc0f54ae2b28fa0c360d5cf00.img

Esempio tabella caratteristiche macchine virtuali

Come già accennato è stata redatta una lista che identifica tutti i file .img appartenenti a ciascuna macchina virtuale.
Questa operazione molto impegnativa ha consentito la migrazione seriale delle macchine virtuali da OVS a Proxmox VE nel più breve tempo possibile.
Le macchine virtuali sono state migrate una alla volta, per la grande dimensione dei file da copiare e soprattutto perché i servizi erogati da un particolare server sono rimasti interrotti per tutta l'operazione.
Innanzi tutto si è spenta la macchina virtuale da migrare in funzione su OVS.
Lo spegnimento è stata una operazione indispensabile per evitare il rischio di avere uno stato difettoso del sistema operativo della macchina virtuale.

Successivamente si sono copiati tutti i file .img associati alla macchina dal nodo OVS che li conteneva al nodo di Proxmox che li avrebbe importati.
Poi sul suddetto nodo di Proxmox è stata creata una macchina virtuale con le stesse caratteristiche fisiche della macchina esportata da OVS, in particolare è stato creato anche un disco (o più) delle dimensioni approssimative di quello esportato da OVS.
Tale disco è stato poi dissociato dalla nuova macchina virtuale ma ne è stato annotato il nome.
Poi è stato convertito il disco importato da OVS su Proxmox dal formato RAW al formato QCOW2. Questa operazione non strettamente necessaria è stata eseguita per fornire una migliore compatibilità col nuovo sistema.
Infine l'operazione più delicata, l'importazione del disco QCOW2 nella nuova macchina virtuale col nome del disco che era stato creato e poi dissociato.
In alcuni casi[5], all'accensione della nuova macchina virtuale (copia di quella esportata da OVS) il sistema operativo Linux ha segnalato la presenza di hardware sconosciuto e questo ha creato un blocco del processo di boot segnalando un errore in initramfs[6]. In questo caso è stato necessario usare il tool dracut per creare un nuovo initramfs corretto.

L'installazione e le opzioni di configurazione di Proxmox usati per questo primo cluster sono molto simili a quelle usate per il sistema finale quindi per ora verrà omessa una loro descrizione accurata rimandando l'argomento al paragrafo relativo al cluster di Proxmox finale.
Verranno però indicati i comandi di importazione di una macchina virtuale su un nodo del cluster Proxmox.

ci si sposta nella cartella di importazione sul nodo dove è stata creata la nuova macchina virtuale
cd /rpool/data/

si importa il file xxxx.img

[5] Per i sistemi operativi Centos 6 e Centos 5
[6] initramfs viene usato in fase di accensione della macchina per caricare il kernel e contiene, fra le altre cose, la descrizione dei device fisici

rsync -P tacchini@moon.hide.bo.iasf:/home/tacchini/xxxx.img .

si converte nel formato qcow2
qemu-img convert -f raw -O qcow2 xxxx.img xxxx.qcow2

si importa nella macchina virtuale
qm importdisk 104 xxxx.qcow2 oasnfs1

Per quel che riguarda l'ultimo comando, il comando di importazione vero e proprio, si può notare che 104 è l'id della macchina virtuale nuova e oasnfs1 è lo storage in cui è stato creato.
Il comando è necessario perché il file system di Proxmox è basato su un database (database-driven) quindi non si può semplicemente copiare.
Terminata l'importazione si può cancellare il file del disco virtuale che si era creato insieme alla nuova macchina virtuale e che non può più essere usato.

PROXMOX VE

Proxmox Virtual Environment è un hypervisor, ovvero un sistema di virtualizzazione, basato su Linux che può essere installato sia su sistemi Debian che bare metal[7].
Con un hypervisor è possibile risparmiare energia e ridurre i costi legati all'IT, inoltre si introduce una grande flessibilità di utilizzo e la possibilità di effettuare backup di macchine intere.
È un software open source quindi non c'è un costo per installarlo ed usarlo, ma c'è per avere supporto e per poter accedere agli aggiornamenti[8].
Un server ospite su cui viene installato Proxmox viene chiamato nodo ed è possibile installarlo in un cluster di più nodi[9] in modo da avere Alta Disponibilità[10].
Le sue caratteristiche principali sono le seguenti:

---

[7] installato direttamente sull'hardware senza avere già un sistema operativo
[8] sono disponibili comunque gli aggiornamenti Linux
[9] almeno 3 nodi
[10] HA High Availability

Macchine virtuali di tipo KVM

La tecnologia **K**ernel-based **V**irtual **M**achine è molto efficiente in quanto opera come un modulo del kernel Linux della macchina ospite.
Questo permette delle performance quasi uguali al sistema puro nel caso si abbiano CPU con supporto nativo alla virtualizzazione tipo le tecnologie Intel VT-x o AMD-V.
Ogni macchina virtuale emula il proprio hardware: CPU, scheda grafica, scheda di rete, ecc.
Le macchine virtuali così ottenute sono delle instance separate le une dalle altre.

Linux Containers (LXC)

Un tipo di virtualizzazione che opera a livello di sistema operativo e non di kernel.
È una tecnologia leggera ed agile in quanto condivide il kernel col sistema ospite.
Usata soprattutto per creare instance legate ad una applicazione.
Per esempio un server web creato in un container è più isolato rispetto al sistema ospite e questo fornisce maggiore sicurezza.

Sistema di Management Centralizzato

Il sistema di management di Proxmox VE è accessibile interamente da interfaccia web, sebbene esista anche l'interfaccia CLI[11], da cui è possibile realizzare tutte le operazioni più importanti.
Per esempio è possibile creare e gestire le macchine virtuali ed i container, migrare le macchine da un nodo ad un altro, definire le attività di alta disponibilità, aggiungere e gestire diversi tipi di storage, aggiornare un nodo ospite, effettuare backup o live snapshot di macchine virtuali.
Realizza un design multi master, significa che è possibile gestire tutto il cluster collegandosi ad uno qualunque dei nodi che lo compongono.

[11] Command Line Interface

Cluster HA[12]

Un cluster di più nodi di Proxmox VE può realizzare una infrastruttura ad alta disponibilità in maniera nativa e semplice da configurare.
È sufficiente indicare nelle caratteristiche descrittive di una macchina virtuale l'opzione di Alta disponibilità desiderata, ad esempio si può migrare la macchina ad un altro nodo in caso di guasto su quello che la ospita.

Networking

Proxmox VE virtualizza l'infrastruttura di rete presentata alle macchine virtuali su ogni nodo.
La simulazione è completa e consente di creare reti virtuali anche molto complesse.
Implementa un modello di networking di tipo bridged, il che significa che vengono realizzati via software degli switch uguali a quelli fisici e per collegarsi al mondo esterno questi vengono associati alle interfacce fisiche del nodo ospite.

Firewall

Viene realizzato un servizio completo di Firewall in modo distribuito tra i nodi, in questo modo è possibile inserire regole a livello di singolo nodo ma anche di singola Macchine Virtuale/ Container.

Backup

Sono supportati in modo nativo sia backup a freddo che snapshot di macchine virtuali e container.
Si possono programmare i backup in modo automatico e gli storage supportati sono di tutti i tipi (anche se per i live snapshot è consigliato usare Ceph).
Esiste anche un prodotto dedicato alla gestione dei backup in generale che è facilmente integrabile con Proxmox VE.

---

[12] High Availability

La funzione di live snapshot è cruciale per quelle macchine mission critical che devono garantire un downtime sotto il 5%.

Storage

La gestione dello storage è molto flessibile, le macchine virtuali possono risiedere sullo storage locale di un nodo oppure su uno storage condiviso, per esempio un NFS visibile da tutti i nodi.
A livello locale viene supportato, oltre che la normale struttura a directory del filesystem, anche LVM[13] e ZFS[14].
L'utilizzo di storage in locale ha la limitazione di non permettere la migrazione in tempo reale delle macchine virtuali, quindi di fatto impedisce l'Alta Disponibilità.
Per sviluppare in pieno il potenziale di questo hypervisor è consigliato usare uno storage condiviso.
Le tecnologie di storage condiviso supportate sono: iSCSI, NFS, SAMBA, Ceph RBD[15]. Ed i file system condivisi GlusterFS e CephFS.
Soprattutto Ceph come file system si è rivelato uno strumento molto potente per gestire i dati delle macchine virtuali.
È un file system distribuito e replicato che sfrutta i dispositivi fisici che stanno sui singoli nodi ma non solo, può utilizzare anche macchine dedicate allo storage e presentarlo a tutti i nodi del cluster.
È flessibile, sicuro e scalabile.

INSTALLAZIONE NUOVO CLUSTER PROXMOX

Sono state acquistate 3 macchine per realizzare un cluster di Proxmox VE che possa garantire Alta Disponibilità e che consenta l'utilizzo di CephFS.
È stato acquistato anche uno switch Netgear XSM4324CS switch 24 porte 10 Gbs per realizzare una sottorete privata di collegamento tra i nodi.

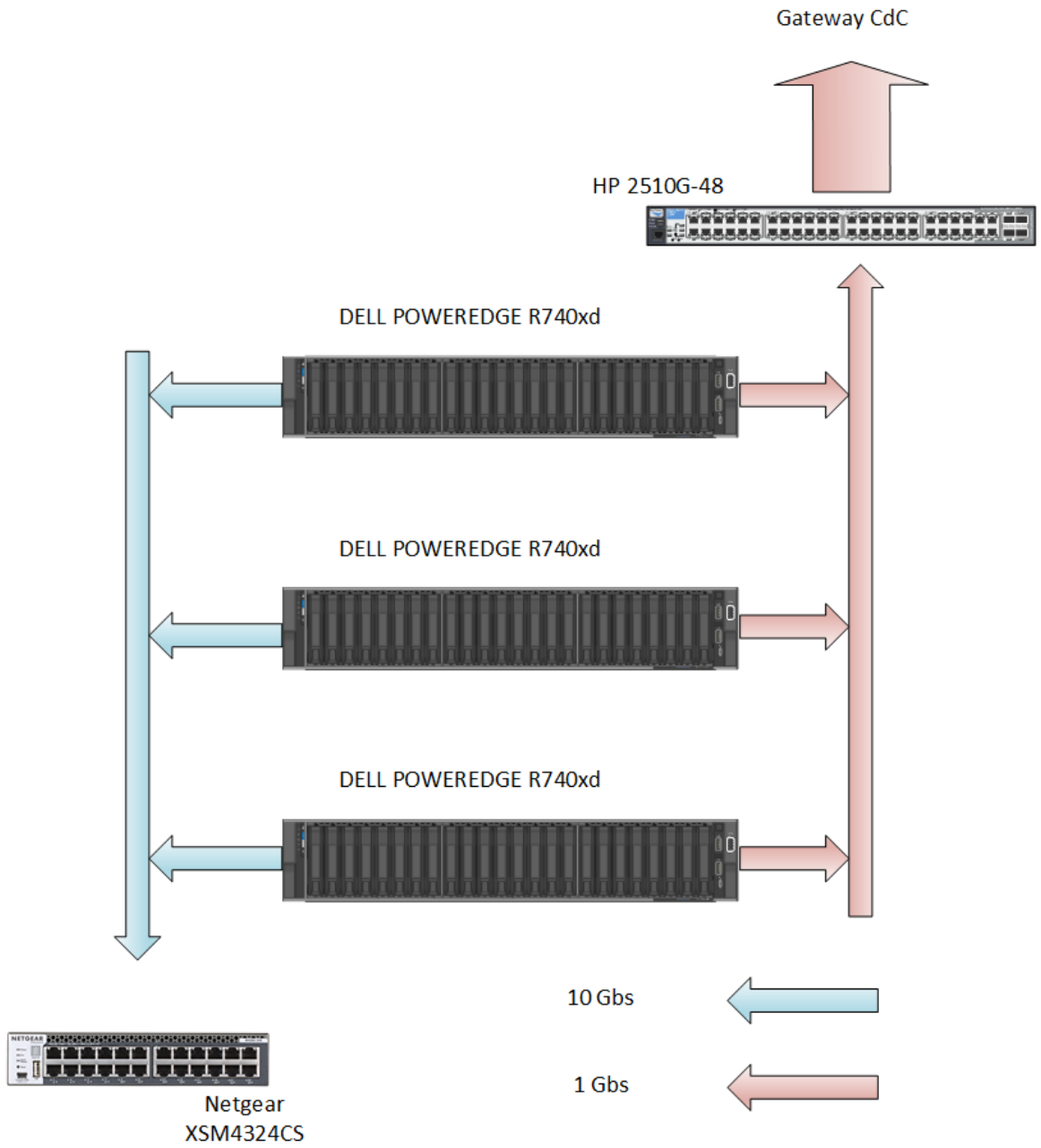
---

[13] Logical Volume Manager
[14] Zettabyte file system
[15] Ceph Rados Block Device

Le caratteristiche per ogni nodo sono le seguenti:

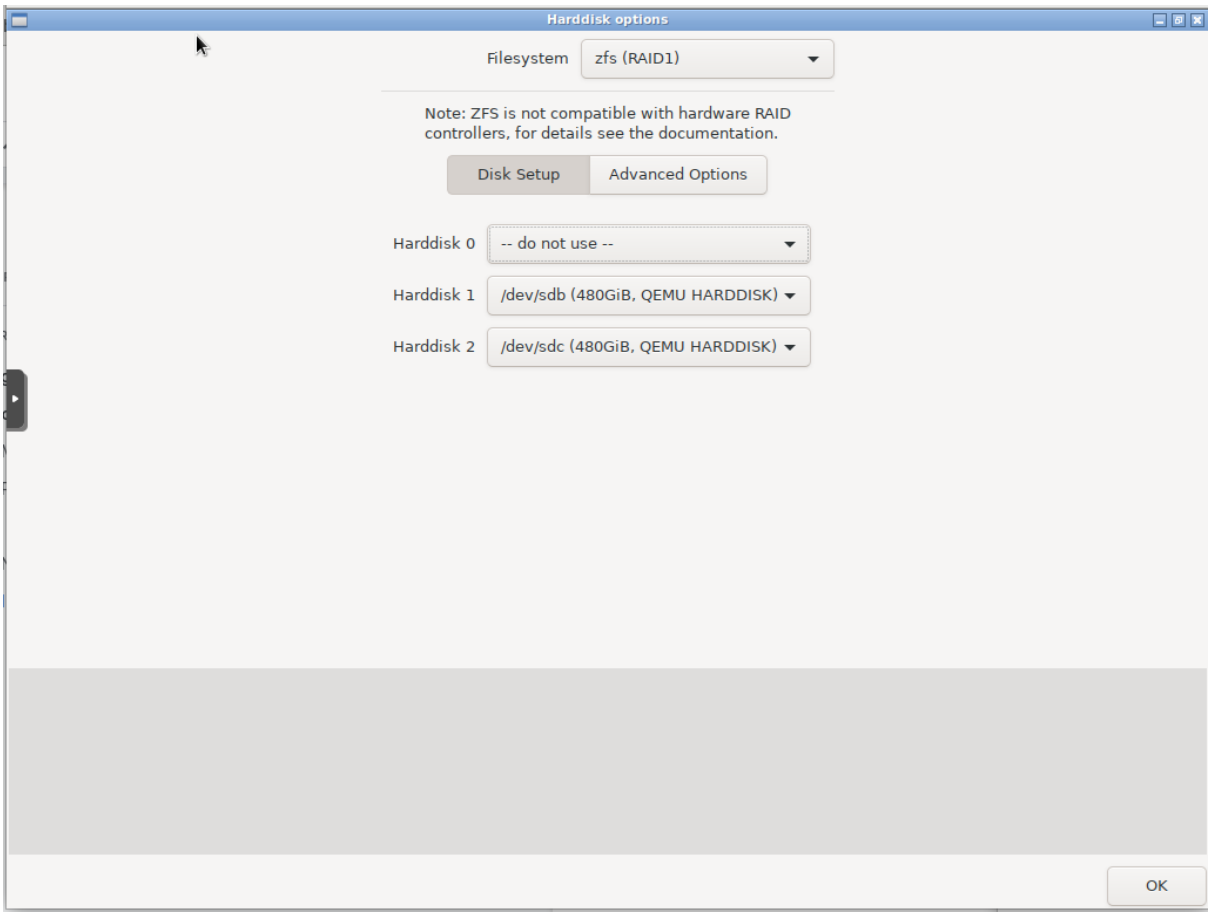
- DELL POWEREDGE R740xd;
- Doppio processore Intel Xeon Gold 6238R 2.2G, 28C/56T, 10.4GT/s, 38.5 M Cache per un totale di 56 Core;
- 128 GB di RAM;
- 2 dischi ssd da ….per il sistema operativo;
- 3 dischi Intel s4610 ssd 1,92TB per realizzare il sistema Ceph;
- Una scheda di rete INTEL X550T2 doppia porta a 10Gbs;

Lo schema seguente raffigura il sistema realizzato:



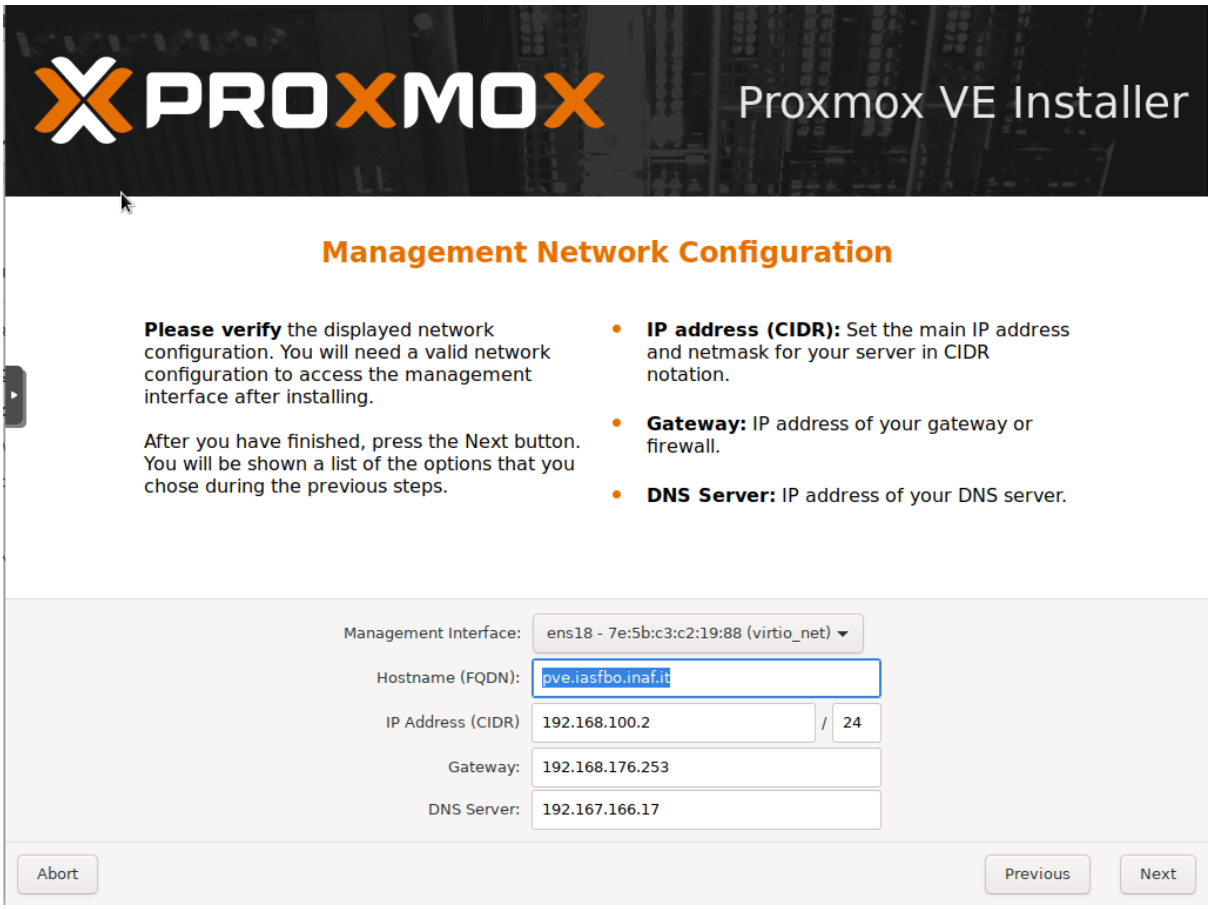
Installazione di Proxmox VE

Il software nella versione 6.4-8 è stato installato mediante una chiavetta usb eseguibile su ognuno dei 3 nodi identici.
Sono stati scelti i due dischi SSD da 480 GB per il sistema ed è stata scelta l'opzione di configurazione zmirror

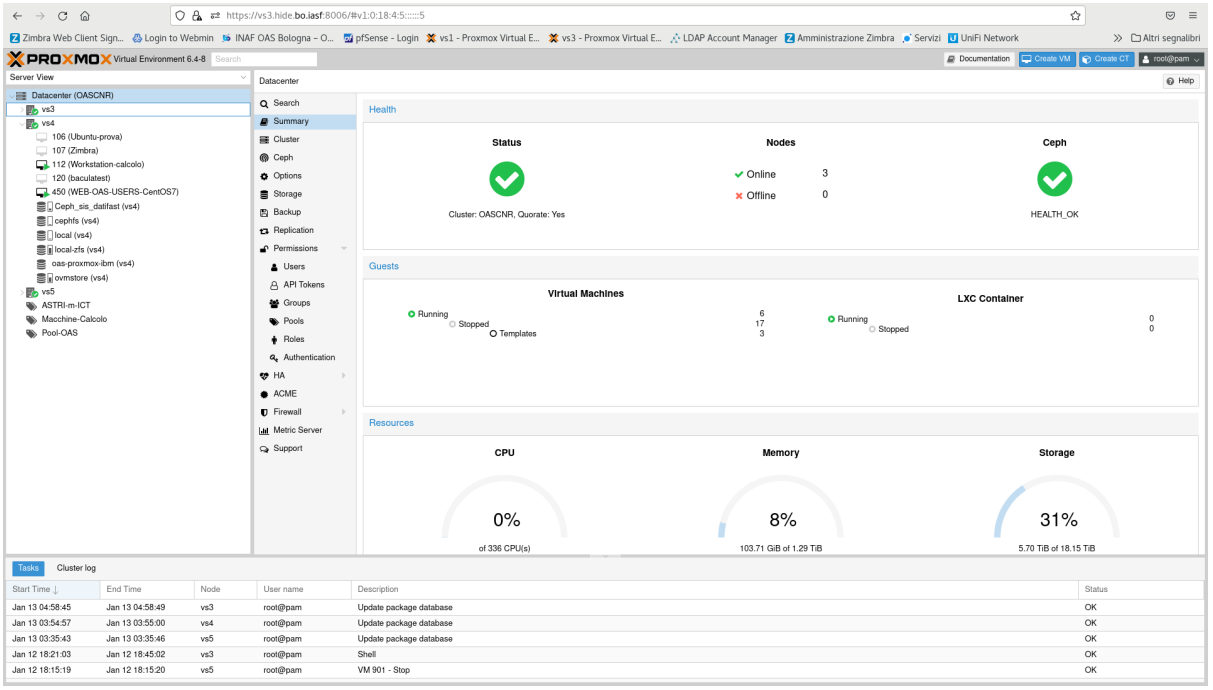


Lo z mirror è un raid 1 creato da ZFS, quindi ha tutte le opzioni disponibili di ZFS (sono state lasciate le impostazioni di default).
Successivamente si sono impostate la lingua, la zona geografica ed il layout di tastiera. Poi la password di amministratore ed un indirizzo email di riferimento, l'email è obbligatoria per procedere.
L'ultimo passo è la configurazione della rete. Sebbene si consigliato di separare la rete che collega verso l'esterno, che usa l'interfaccia in modalità bridge, la rete che realizza il cluster (usa Corosync e necessita di bassa latenza) e la rete che collega lo storage (nel nostro caso Ceph), si sono lasciate le prime due su un unico collegamento fisico perchè

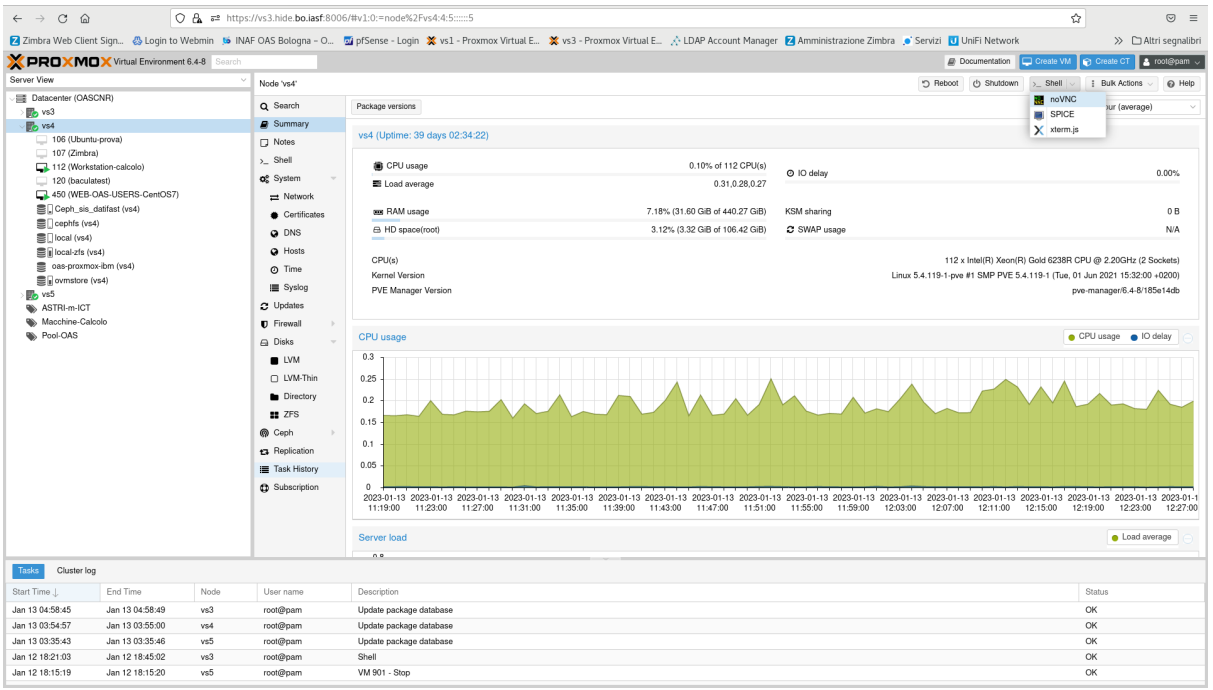
tutto sommato Corosync richiede un collegamento stabile ma poca banda e viene utilizzato uno switch fisico dedicato a questi collegamenti. Questo switch è collegato poi al centro stella del Centro di Calcolo.



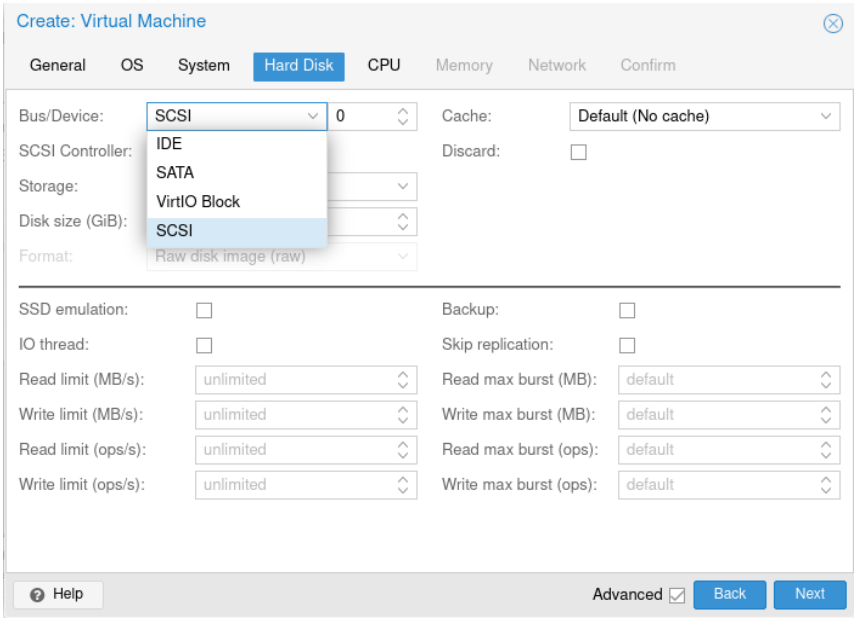
Terminata l’installazione su ogni nodo si può accedere, mediante browser, all’interfaccia grafica di controllo.



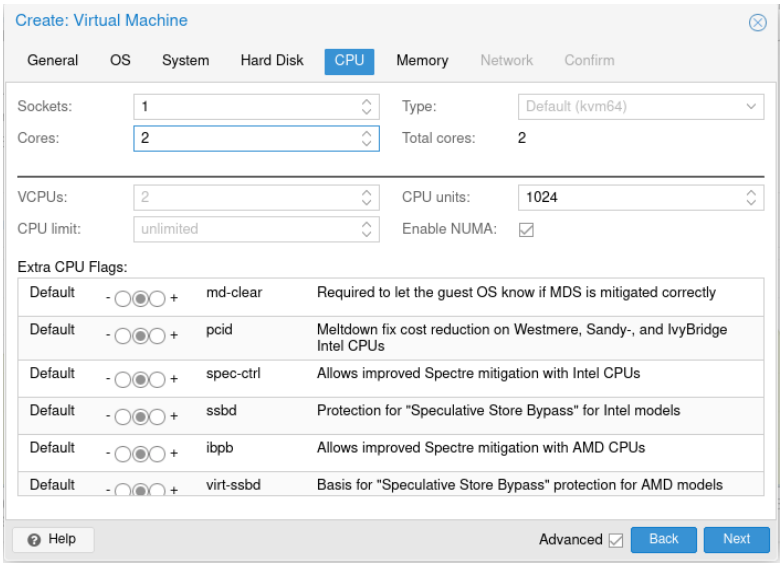
In fase di installazione ci si è assicurati di abilitare la porta seriale in modo da avere disponibile anche xterm.js come console (può essere utile per il copia/incolla dei comandi).



Per la creazione di una macchina virtuale ci sono alcune accortezze da tenere presente. Per lo storage normalmente si usa SCSI o VirtIO Block, se si dovesse installare una macchina Windows è opportuno impostare IDE (per via dell'installer di Windows) e poi eventualmente cambiare dopo. Si può scegliere già in questa fase se effettuare un backup regolare del disco.

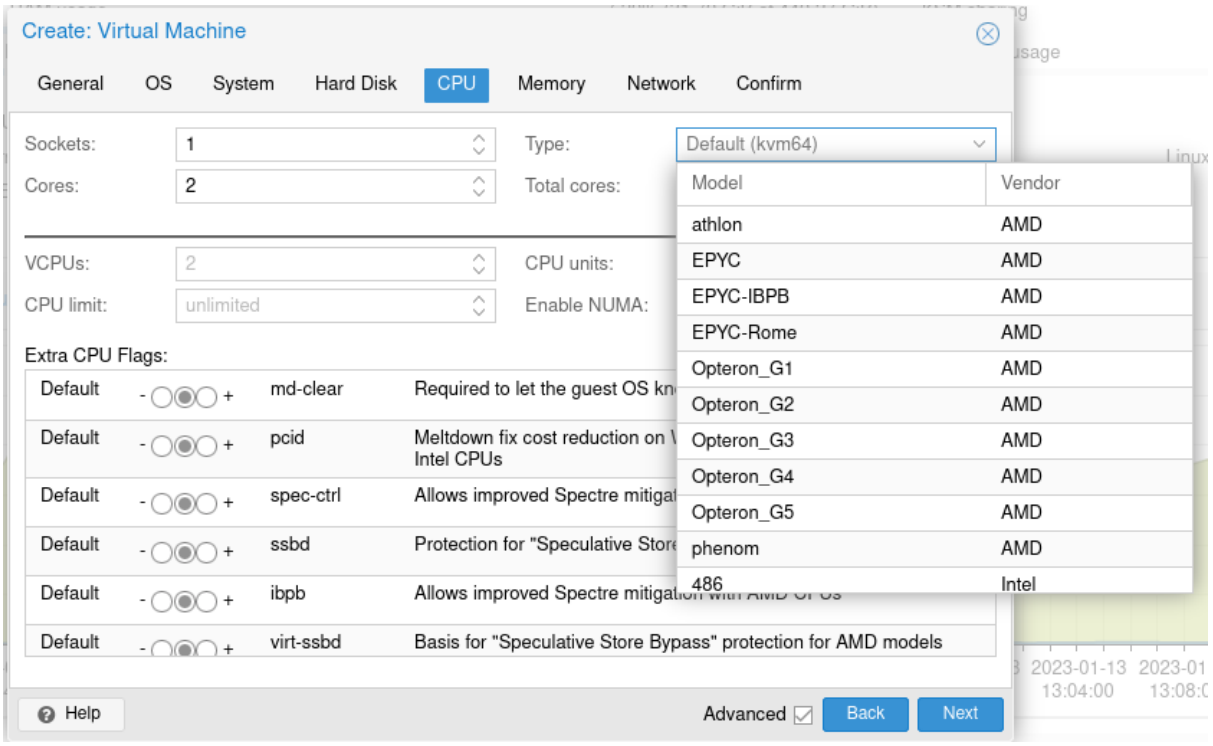


Per quel che riguarda la CPU andrebbero sempre scelti almeno due Core e tenuta sempre abilitata l'opzione NUMA[16]

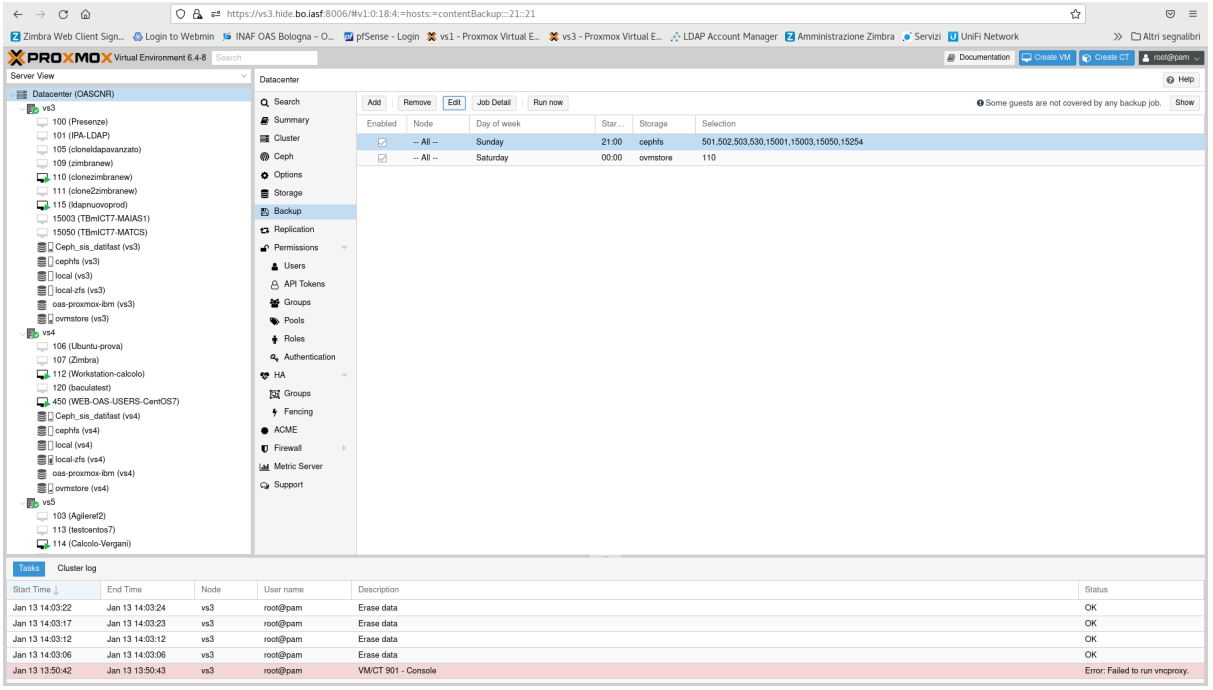


[16] Non Uniform Memory Access

Nel caso ci siano problemi di installazione, per sistemi operativi ostici o perché si ha un cluster Proxmox molto disomogeneo, andrebbe settato il campo Type con la famiglia di processori fisici presenti sul nodo di Proxmox.

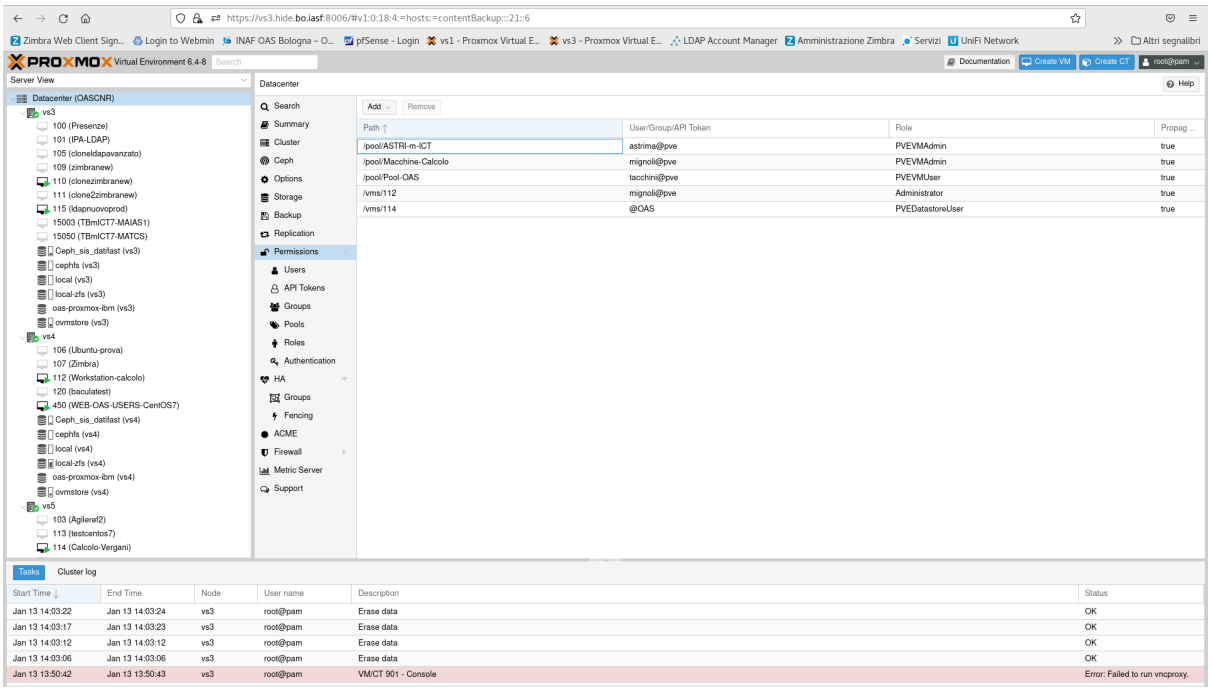


Il backup si può effettuare manualmente, usando una delle tre modalità: macchina ferma, snapshot, o macchina in sospensione; oppure si può schedulare attraverso l'apposita voce in Datacenter (le risorse/attività configurate in Datacenter riguardano il cluster intero).



La configurazione del backup è abbastanza complicata da riassumere ma, per brevità, è sufficiente prestare attenzione allo storage che verrà usato (per evitare di saturarlo) ed alla modalità di backup: *snapshot* per macchine che devono restare accese e *stop* per macchine che possono essere spente e per cui l'integrità del backup è importante.

Proxmox consente una gestione dei permessi molto estesa, si articola su utenti, gruppi, e sui ruoli che si possono avere sui pool di risorse (come lo storage o le macchine virtuali intere).
Si configura sempre da Datacenter.



L'alta disponibilità su Proxmox (HA[17]) è basata su Corosync e richiede almeno 3 nodi per funzionare. Il quorum per un cluster di 3 nodi come il nostro è di 2.
C'è un meccanismo di fencing integrato (Softdog Kernel o Watchdog Kernel) per gestire i guasti ai nodi ma è possibile usare anche meccanismi esterni.
Anche l'HA si configura da Datacenter.

Ceph

Ceph è un file system condiviso sviluppato da Red Hat ed è opensource. Questo sistema è costituito da uno storage distribuito a oggetti che mette a disposizione degli utenti un object storage, un block storage (RBD) ed un File System Posix.

[17] High Availability

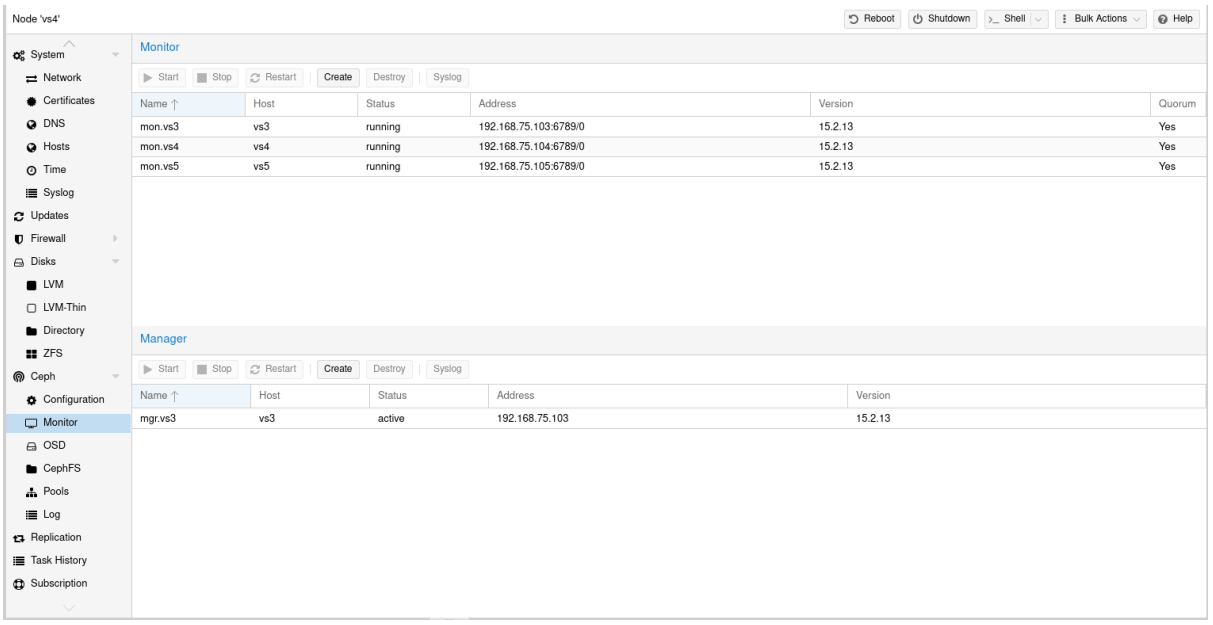
Ceph è dotato dai seguenti componenti:

- Monitor: per il controllo del cluster. Necessita di almeno 2GB di RAM. Ogni nodo è stato settato come monitor.
- OSD (Object Storage Device): sono i singoli hard disk. Per ottimizzare le prestazioni si è deciso di usare dischi SSD di classe datacenter e tutti uguali.
- Manager: insieme ai monitor monitorizza e gestisce il cluster. Nel nostro sistema è vs3-
- MDS (Meta Data Server): contiene i metadati di CephFS.
- Gateway: una interfaccia l'object storage e le applicazioni.

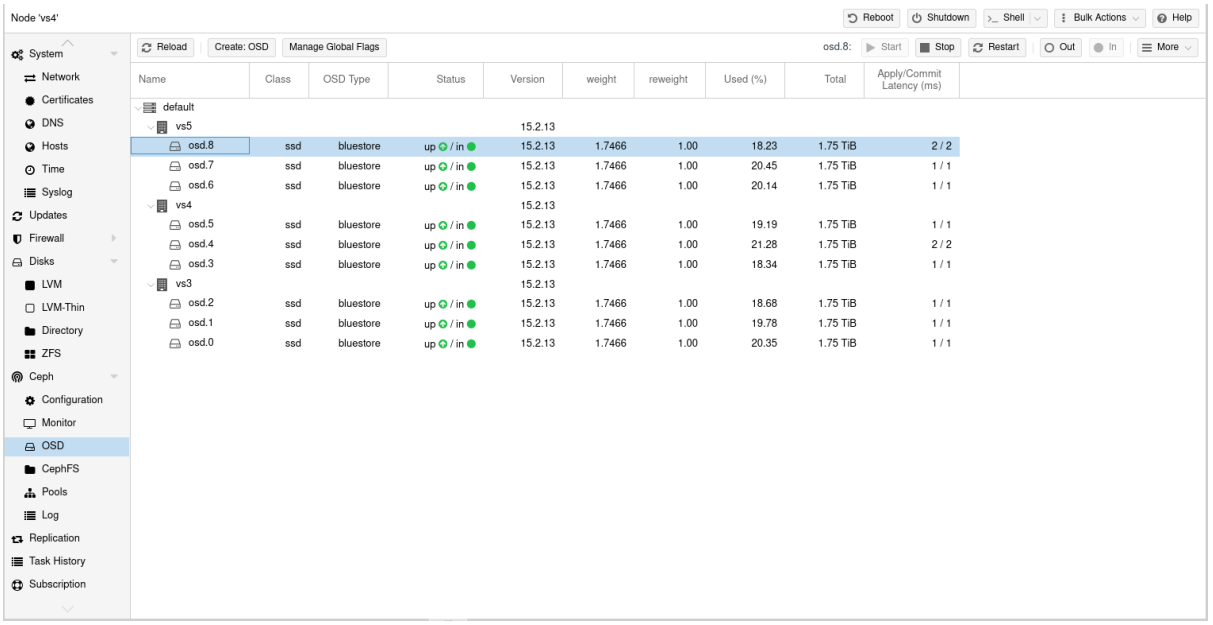
Alcune proprietà importanti di cui è dotato intrinsecamente Ceph sono lo Thin Provisioning, la Compressione dei dati, Snapshot, Copy on Write, Load Balancing adattivo, No Single Point of Failure (solamente con 3 o più nodi).
Prima di installare Cep si sono seguite alcune raccomandazioni.
Avere un cluster di almeno 3 nodi per realizzare la No Single Point of Failure.
Avere una sottorete dedicata ai dati realizzata con uno switch fisico a 10Gbs.
Usare dischi veloci evitando i dischi meccanici.
Dimensionare la RAM di ogni nodo in modo da soddisfare il requisito di 3-5 GB di RAM per TB (nel nostro caso significa almeno 30 GB di RAM a nodo).
Dimensionare la CPU in base alla formula: 2 GHz per ssd / velocità in GHz del singolo Core.
Non usare RAID ma presentare direttamente i dischi a Ceph.

Per l'installazione vera e propria si procede per ogni nodo perché non c'è un controllo centralizzato su Datacenter.
Si può procedere via CLI oppure via interfaccia grafica.
Si è scelta quest'ultima modalità.
Dalla voce Ceph si sceglie Start Installation, si sceglie la modalità advanced e si prosegue inserendo la subnet dedicata per Ceph ed impostando il nodo stesso come monitor.

Si arriva alla situazione seguente:



A questo punto vanno inseriti gli osd per ogni nodo.
Scegliere Ceph - - OSD - - Create OSD.
Scegliere il disco e Bluestore per OSD type.
La situazione finale:



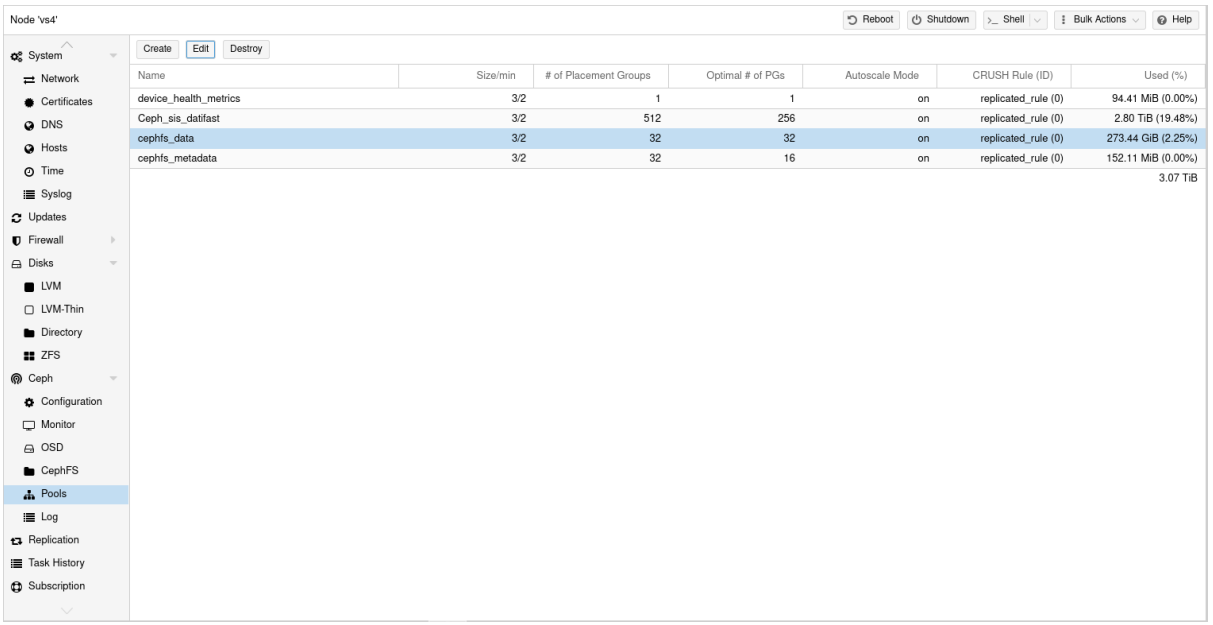
Potrebbe essere necessario reinizializzare dei dischi già usati.
Si usano i comandi:

dd if=/dev/zero of=/dev/sdx bs=1M count=200
ceph-zap /dev/sdx

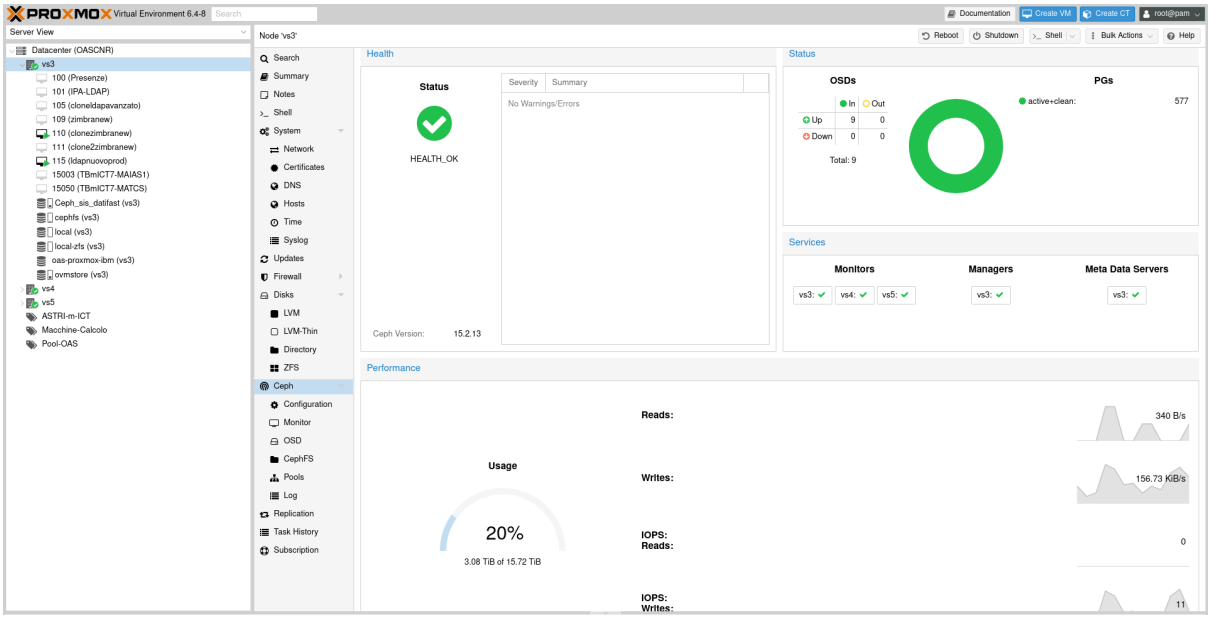
Per avere lo storage disponibile su nodi è necessario creare un pool
Da un qualsiasi nodo scegliere Ceph - - Pools - - Create, impostare il nome del pool, settare la size max e la size min (in pratica size max andrebbe settata pari al numero dei odei mentre la size min è la soglia minima di funzionamento, per 3 nodi andrebbe messa a 1 ma si è scelto comunque 2 per non rischiare perdita di dati e perché, data la struttura del cluster, è improbabile un malfunzionamento), sceglie la crush rule (c'è solo un'opzione), impostare il numero di Placement Group (PG) (in genere il valore è 32 ma viene messo a disposizione un calcolatore online https://ceph.com/pgcalc), infine con Add Storage si aggiunge il pool come storage in Datacenter (è da Datacenter che si gestisce qualunque storage).
Il PG serve a ridurre la quantità di metadati associata agli oggetti sostanzialmente raggruppandoli in gruppi, è necessario un bilanciamento in quanto un alto numero di PG migliora la granularità ma aumenta il peso computazionale.
Si è deciso di sfruttare Ceph sia come File System che come RBD[18].
Il File System è stato chiamato CephFS e, per semplificare, viene utilizzato come un NFS molto performante (molto veloce e moto robusto), opera a livello di file.

[18] Radon Block Device

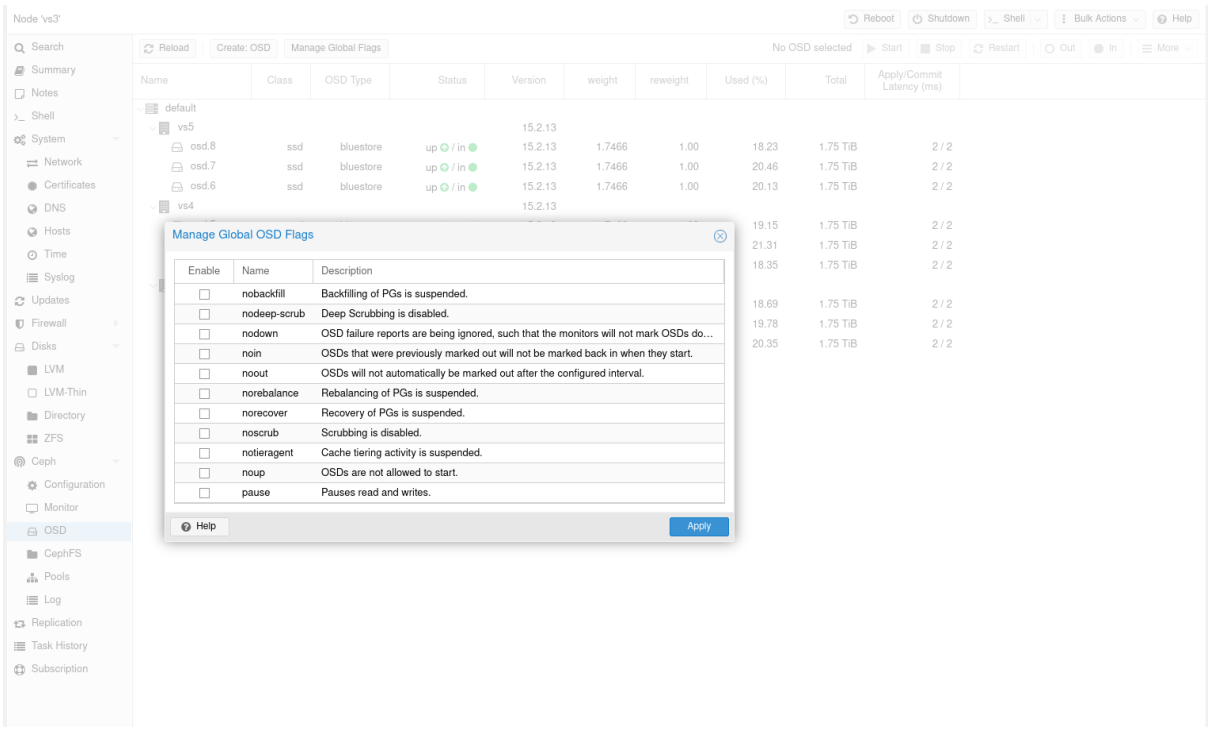
Lo storage RBD è stato chiamato Ceph_sis_datifast ed anch'esso è condiviso tra tutti i nodi ma a livelli di Block Device, viene utilizzato prevalentemente per i dischi delle macchine virtuali.



Per vedere lo stato di Ceph:



Una nota importante è che se si deve intervenire su uno degli OSD (per esempio per sostituirlo in casi di guasto) oppure se si deve riavviare un nodo è necessario settare il/gli OSD con il flag noout per evitare rischi di malfunzionamenti.



Per concludere un’ultima osservazione, è risultato fondamentale aver pianificato in anticipo sia l'obiettivo di funzionamento che il dimensionamento dell’hardware.
Ha semplificato molto la messa in opera di tutto il sistema.