High Availability con VMware Server (free)

Descrizione:

Dato il continuo aumento delle nuove macchine da installare anche noi abbiamo pensato ad un sistema di virtualizzazione.

Abbiamo scelto VMware per vari motivi tra cui l'esperienza fatta con questo strumento, per la sua semplicita' di gestione e per la possibilita' di eseguire macchine virtuali Windows non avendo processori con virtualizzazione hardware.

Inizialmente e' stato testato VMware ESX con grande soddisfazione al punto da interessarci all'acquisto. ESX ha la particolarita' di funzionare in cluster su un disco condiviso iSCSI o FC e permette di spostare le macchine virtuali praticamente senza alcun interruzione, in caso di crash di una macchina fisica e' in grado di "resuscitare" le VM su un altro nodo.

Purtroppo il costo della versione ESX appare praticamente irraggiungibile ed abbiamo desistito, ma abbiamo pensato a come mettere su una struttura simile ultilizzando la versione free VMware Server.

La soluzione e' stata la gestione delle macchine virtuali di VMware Server tramite un RedHat Cluster Suite (RHCS) su Scientific Linux 4 con storage condiviso iSCSI / Fiber Channel.

Lo storage viene gestito tramite il Clustered Logical Volume Manager di RHCS, inserendo volumi esportati da sistemi SAN come Phisical Volume LVM in un Volume Group, questo viene poi partizionato creando un Logical Volume per ogni macchina virtuale, dove risiederanno i file di configurazione e i dischi virtuali VMWare.

Tramite il cluster siamo riusciti ad ottenere l'affidabilita' della versione ESX di VMWare nonche' la possibilita' di migrare da un host all'altro le virtual machine con una breve interruzione (max. 20 secondi) senza fare lo shutdown del sistema guest.

Sono stati preparati degli script init-like per far gestire a RGManager (il gestore dei servizi di RHCS) le virtual machine come dei servizi del cluster, gli script supportano "start | stop | status" che sono implementati nel seguente modo:

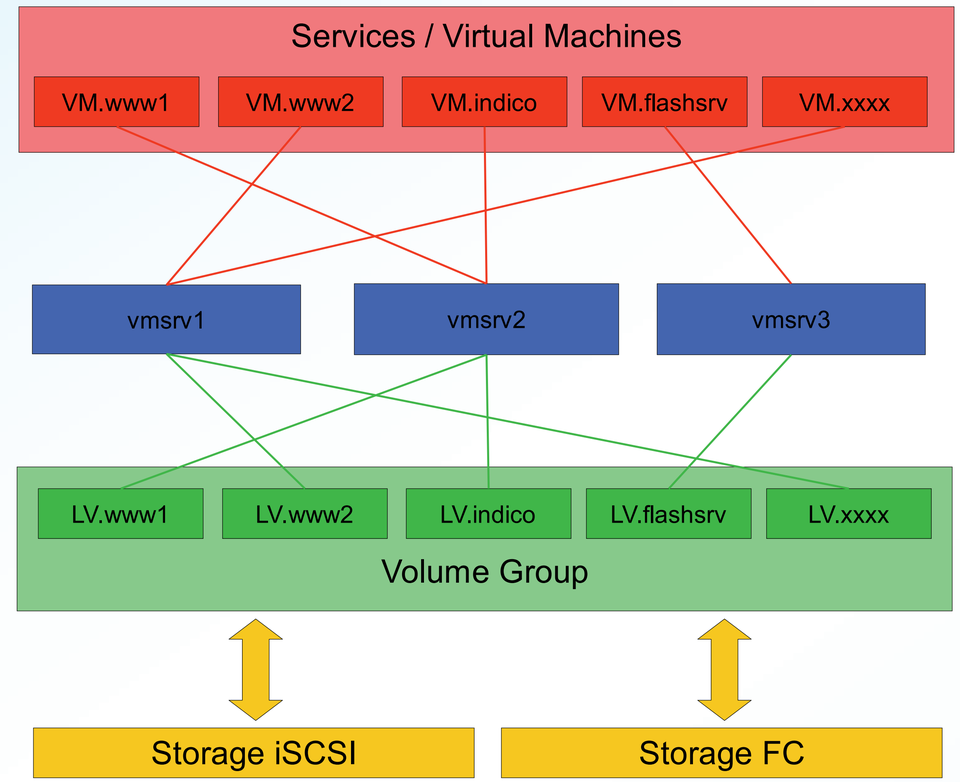
Start: Registra la macchina virtuale in VMware ed esegue lo start, risponde inoltre di mantenere l'UUID della Virtual Machine

Stop: Esegue il suspend delle macchina virtuale e rimuore la registrazione in VMware

Status: Chiede a VMWare lo stato della macchina virtuale e risponde con exit-code di conseguenza

RGManager controlla periodicamente lo stato dei servizi, se un servizio non dovesse rispondere tenta di riavviarlo prima sullo stesso host poi su un altro.

Se una macchina fisica dovesse andare in crash le virtual machine presenti su di essa verranno automaticamente riavviate su altri nodi del cluster. In tal caso il sistema operativo guest deve necessariamente eseguire di nuovo il boot non avendo salvato preventivamente il proprio stato.



Cluster

La disposizione delle macchine virtuali viene amministrata tramite RGManager il quale e' in grado di far partire, di controllare lo stato e fermare i servizi nonche' creare dipendenze tra un servizio e uno storage.

Sono stati creati appositi script init-like per la gestione delle VM di VMWare.

Se un nodo va in crash le VM verranno ripristinate su un altro nodo, dove RGManager montera' il filesystem necessario.

Lo stop di una VM e' in realta' un suspend della stessa, in questo modo il relocation manuale di una VM comporta un'interruzione di circa 20 secondi senza che questa debba rieseguire il boot.

Host

Blade HP BL25p G1, 2 x dual Opteron 280 64bit @ 2.4 GHz, 11GB RAM, 4 x Ethernet, 2 x FC, switch GbE e FC nello chassis. OS: Scientific Linux 4.4 x86_64.

Le interfacce Eth sono configurate in HA (active-backup) tramite linux bonding

LVM

Vengono creati volumi per ogni macchina virtuale grazie al Clustered Logical Volume Manager.

I Logical Volume possono essere creati, estesi ed eliminati da qualsiasi nodo del cluster e l'informazione e' automaticamente propagata agli altri nodi.

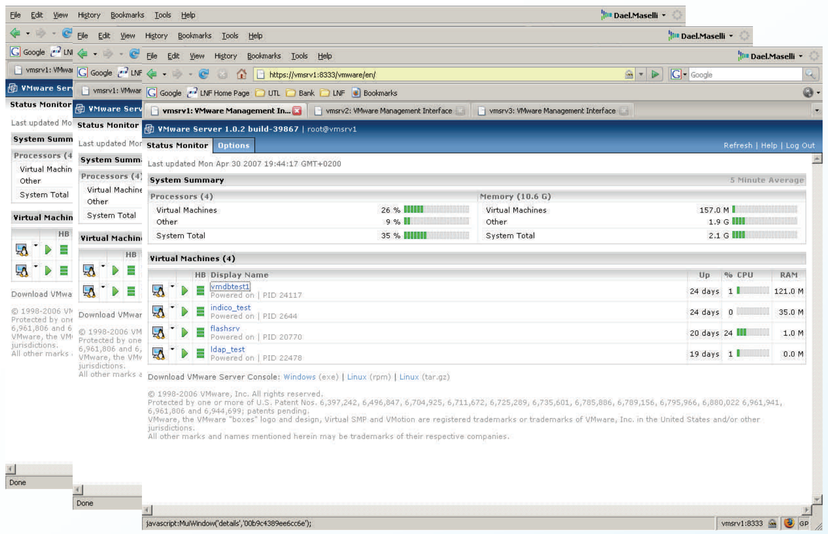
Il file system sui LV e' ext3, tramite resize2fs puo' essere ridimenzionato on-line.

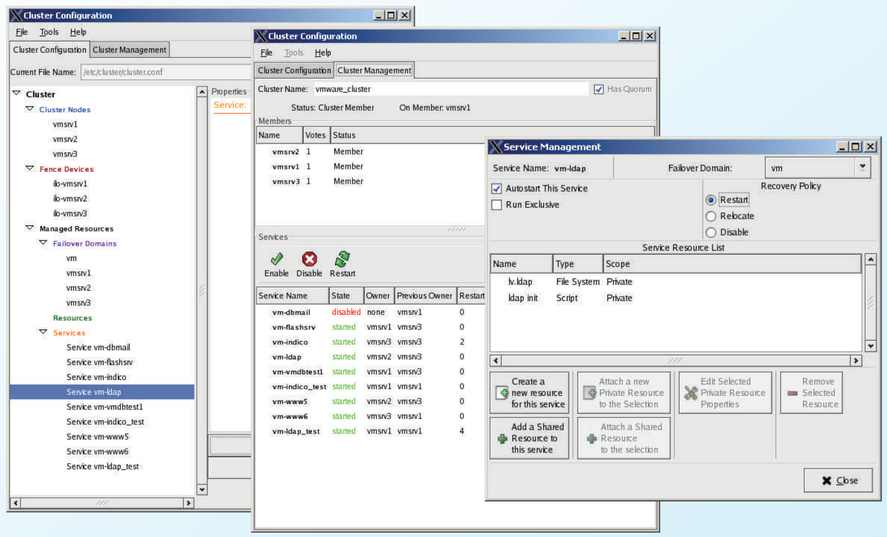
E' compito di RGManager occuparsi di montare i volumi secondo le dipendenze dei servizi ed assicurarsi che non siano montati su altri nodi.

SAN

Lo storage puo' risiedere su un qualsiasi hardware che permetta l'accesso condiviso ai volumi raw.

E possibile inserire piu' tipi di storage all'interno dello stesso Volume Group ed eventualmente evacuare un Phisical Volume per effettuare delle migrazioni ad esempio da iSCSI a FC o viceversa





- Dael Maselli -

LNF - INFN Computing Service, Via Enrico Fermi, 40 - 00044 Frascati (RM)

Template by Nunzio Amanzi - LNF INFN Computing Services - April 2007