



"Entry level cluster" con Linux

Una soluzione per realizzare sistemi di tipo "cluster" a costo contenuto

L'esigenza di eseguire applicazioni critiche su sistemi configurati per garantire **alta affidabilità** e **alta disponibilità** è in continua crescita: il costo del disservizio imputabile al mancato funzionamento di un'applicazione, anche per poche ore soltanto, è spesso superiore al costo di una soluzione di *clustering*.

10 anni di alta affidabilità

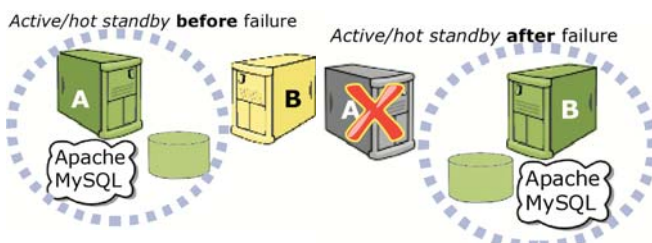
Negli ultimi 10 anni, per eseguire le applicazioni critiche tipiche dell'ambiente "distribuito" (UNIX e Microsoft Windows), sono stati installati sistemi con una o più delle seguenti caratteristiche:

- dischi con ridondanza (RAID 1 e RAID 5)
- hardware ridondato (alimentatori, schede di rete, ecc...)
- accesso ai dischi con ridondanza (*storage area network* con *multipath*)
- software e middleware intrinsecamente ridondato (*database* e/o *application server* multinodo)
- ecc...

Standard de facto

Il sistema ad alta affidabilità che ha riscontrato la maggior diffusione è quello definito come "active/hot standby cluster", costituito dai seguenti elementi:

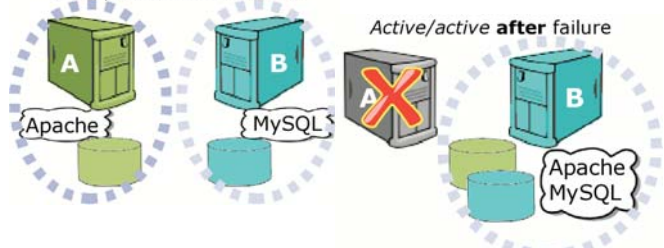
- il cluster si compone di 2 soli nodi
- ciascun nodo possiede un disco locale per il solo sistema operativo
- ciascun nodo è collegato ad uno o più dischi contenenti i dati delle applicazioni attraverso la tecnologia SCSI o, più frequentemente, la tecnologia "fibre channel"; i dispositivi utilizzati sono normalmente in configurazione RAID 1 o RAID 5
- i dischi contenenti i dati delle applicazioni possono essere utilizzati, a turno, da ciascuno dei due nodi
- le applicazioni sono eseguite tutte sul nodo attivo e, nel caso di guasto hardware/software, il *cluster manager* provvede a: arrestare le applicazioni sul nodo guasto, smontare i dischi con i dati dal nodo guasto, montare i dischi con i dati sul nodo secondario, avviare le applicazioni sul nodo secondario.



La soluzione permette di ottenere un sistema che presenta un rapporto fra costo totale (*hardware* e *software*) e disponibilità (tempo di ripristino automatico del servizio) ideale solo per sistemi altamente critici.

Una variante talvolta utilizzata è la soluzione "active/active cluster" per cui, in regime di normale funzionamento, le applicazioni sono divise in modo statico fra i due nodi e, in caso di guasto, il nodo "sano" rileva i servizi normalmente assegnati al nodo guasto.

Active/active before failure



Questo secondo tipo di soluzione permette di sfruttare le prestazioni dell'hardware di entrambi i nodi durante il funzionamento normale ma comporta una diminuzione delle prestazioni in caso di emergenza.

Il problema

La realizzazione di un cluster in configurazione "active/hot standby" o "active/active" presenta dei costi di allestimento ed esercizio superiori rispetto ad un sistema tradizionale ("single node") senza soluzioni di "alta affidabilità"/"alta disponibilità":

- la presenza di due nodi attivi, con relativo sistema operativo, comporta un raddoppio dei costi di acquisizione e manutenzione del sistema operativo
- alcuni software proprietari per la gestione del cluster sono molto costosi sia in termini di licenza che in termini di manutenzione
- alcuni software proprietari (*database*, *transaction manager*, ecc...) richiedono l'acquisto di una licenza doppia anche nel caso in cui il servizio, in un certo istante, sia attivo solamente su un sistema
- i dischi condivisi da due o più sistemi (SAN - *storage area network*) sono molto più costosi dei dischi di identica capacità dedicati ad un solo sistema e possono rappresentare fino al 60% del costo complessivo dell'hardware.

Poter realizzare una soluzione di cluster "active/hot standby" o "active/active" a costi inferiori permetterebbe di:

- diminuire i costi per le applicazioni che già utilizzano sistemi ad alta affidabilità
- installare in configurazione di alta affidabilità un maggior numero di applicazioni senza dover aumentare il budget.



La soluzione

Il sistema operativo GNU/Linux permette di ridurre sensibilmente i costi di realizzazione di una soluzione ad elevata disponibilità ed affidabilità:

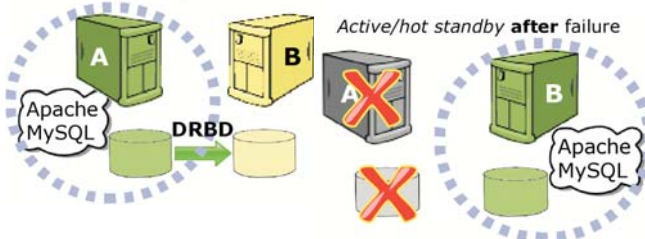
- il costo di licenza del sistema operativo è nullo, il costo annuo per la manutenzione è inferiore rispetto a sistemi proprietari equivalenti
- il software per la gestione del cluster è nullo, in quanto già compreso nelle principali distribuzioni (*High-Availability Linux*)
- il costo di eventuali software proprietari è invariante
- **il costo dei dischi condivisi può essere eliminato utilizzando la tecnologia DRBD.**

La tecnologia DRBD

Definito "dispositivo" una partizione primaria o logica di un *hard disk*, DRBD (*Distributed Replicated Block Device*) permette di effettuare il *mirroring* dei dispositivi in rete anziché localmente: i dispositivi di ciascun sistema vengono accoppiati con quelli corrispondenti del sistema gemello; per ciascuna coppia di dispositivi si assegna il ruolo di *primary* e di *secondary* ai due sistemi: tutte le operazioni di scrittura effettuate sul *primary* vengono riportate sul *secondary* dalla tecnologia DRBD in modo completamente trasparente alle applicazioni.

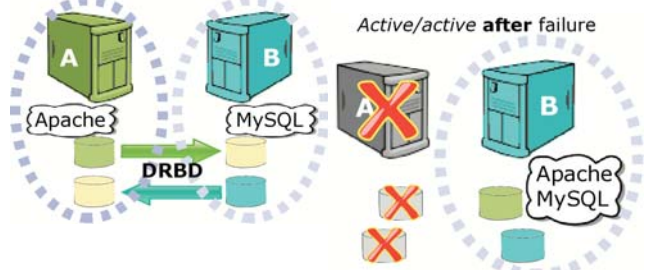
In caso di guasto ad un nodo del cluster, il *cluster manager* provvede a porre *offline* il dispositivo del nodo guasto ed eventualmente a promuovere *primary* quello del nodo sano.

Active/hot standby before failure



La **sincronizzazione** da *master* a *slave* può essere **sincrona** (per privilegiare l'integrità dei dati) o **asincrona** (per privilegiare la velocità di scrittura).

Active/active before failure



L'*hardware* aggiuntivo consigliato è composto da 2 schede di rete da 1Gbit e un cavo cross di qualità: componenti molto più economici dei dischi condivisi in tecnologia SCSI o *fibra channel*. I dispositivi accoppiati non necessitano di ridondanza (RAID *hardware* o *software*) in quanto il sistema complessivo è **intrinsecamente ridondato**. I dispositivi DRBD visti dalle applicazioni sono *standard* in quanto perfettamente aderenti al paradigma VFS (*Virtual File System*).

Applicabilità della soluzione

La soluzione trova la sua naturale applicazione in tutte quelle situazioni in cui le elevate prestazioni fornite da una SAN realizzata con tecnologia *fibra channel* non sono necessarie, ma i prezzi della stessa costituiscono un vincolo al miglioramento della disponibilità e della affidabilità dei sistemi utilizzati per eseguire le applicazioni.

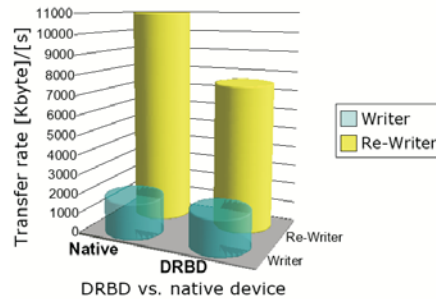
Prestazioni¹

Le misure effettuate nei laboratori PRIMEUR con il *tool* di benchmark "iozone" hanno mostrato che la soluzione DRBD presenta uno *slow down* medio, rispetto al dispositivo fisico nativo di circa il 15% in fase di scrittura e nullo in fase di lettura. Il degrado di prestazioni ottenuto è pertanto più che accettabile per una soluzione che si propone come "entry level".

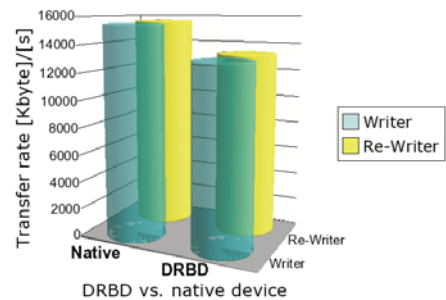
Slow down misurato nei test di laboratorio

Tipo Benchmark	Scritture sincrone	Scritture asincrone
"writer"	1.41%	15.15%
"re-writer"	31.87%	14.76%

IOzone write sync benchmark



IOzone write async benchmark



Elementi chiave della soluzione

Le tecnologie DRBD (*Distributed Replicated Block Device*) e HA (*High Availability*), proprie di Linux 2.6.x permettono di realizzare cluster di tipo "active/hot standby" e "active/active":

- senza utilizzare software proprietario specifico
- senza utilizzare storage condiviso

Consentono pertanto di:

- **aumentare il numero di sistemi** in configurazione cluster della server farm senza aumentare il budget di spesa
- **diminuire il "costo per nodo" dei sistemi** in configurazione cluster della server farm

¹ Le misure sono state effettuate con sistemi di classe pentium III, dischi EIDE e connessione ethernet (cavo incrociato) configurata a 100Mbit/s full duplex utilizzando il software di benchmark "iozone"



PRIMEUR
www.primeur.com



Per ulteriori informazioni
contattare:

sales@primeur.com