SAN@MiB

INFN - Sezione di Milano-Bicocca

Paolo Dini – Servizio calcolo Sez. Milano Bicocca

L'idea che sta alla base di questo progetto e' quella di fornire alla Sezione INFN e al Dipartimento di Fisica di Milano-Bicocca un cluster ad alta disponibilita' per il “public login”, con servizi bilanciati su piu' server e dotato di un sistema di storage ridondato.

(“Toglietemi tutto, tranne i miei mail...”)

La configurazione

Il sistema descritto e' un cluster **HA** di server Linux (SL 4.7 e SL 5.3) connessi tramite tecnologia **Fibre Channel** ad un pool di storage FC in una vera e propria **Storage Area Network (SAN)**. Un primo file system **GPFS** (**/system**) permette la condivisione tra i server dell'area di spool dei mail e delle configurazioni dei servizi offerti dal cluster, oltre che delle eventuali immagini di server virtualizzati (XEN). Un secondo file system GPFS (**/user**) offre spazio su disco agli utenti ed agli esperimenti. La realizzazione di questa SAN si basa largamente sull'esperienza fatta per il cluster di calcolo di CMS (http://cmscluster.mib.infn.it), ma con la sostanziale differenza che in questo caso i file system sono ridondati (data e metadata) su due differenti unita' di storage in maniera da poter essere on line anche con una sola unita' attiva.

Un sistema di Load Balancing posti in testa al cluster permette la distribuzione del carico dei servizi centrali ai quali si accede tramite uno (o piu') virtual server: lettura mail, interattivo, web, spazio su disco (ed eventualmente anche DNS e LDAP, servizi di stampa, ecc.). L'autenticazione avviene tramite i server **LDAP** di sezione.Per la sottomissione di job in batch sono stati installati sia **CONDOR** che **PBS/TORQUE**.

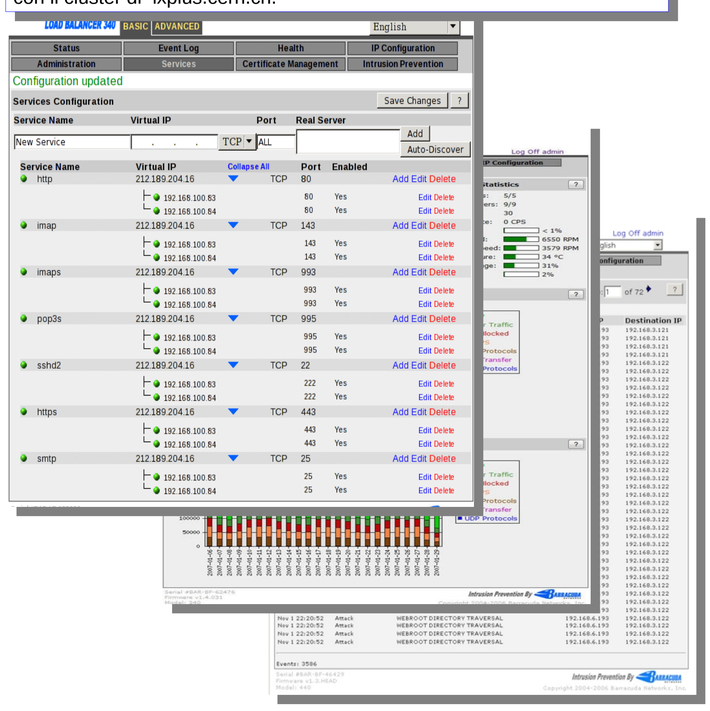
• **POPS/IMAPS**
E' stato scelto DOVECOT per i servizi di lettura della posta, con le INBOX condivise dai server su /system.
•**SMTP**
il servizio e' gestito da POSTFIX, con spool su filesystem condiviso e queue directory locale per ogni server. Le account locali di sistema (root, ecc.) hanno spool locali ridefinite tramite gli alias. I processi procmail locali ai server consentono l'uso dei filtri anti-spam.
•**SSHD2**
•E' un servizio sshd che risponde sulla porta 222 dei real server, con configurazioni e chiavi condivise su /system. Il sistema di Load balancer provvede a fare il forwarding della porta 222 dei real server sulla usuale porta 22 del virtual server . Due sono i motivi di questa scelta: evitare che la rotazione dei real server che rispondono alla connessione si traduca in una nuova richiesta di scambio delle chiavi pubbliche, e la possibilita' di metter in atto una diversa politica di sicurezza rispetto al servizio sshd standard, anch'esso attivo sui real server ma limitato alle connessioni provenienti dalla LAN e dai siti istituzionali.
•**HTTP e HTTPS**
Anche questi servizi sono bilanciati: configurazioni, script, certificati e root directory sono condivisi su **/system**. Su https e' attivo il servizio di lettura mail via web (**squirrelmail**).
•**XEN**
•Potrebbe essere utile trasferire sul cluster alcuni servizi che sono gia' stati virtualizzati con XEN. Le ultime versioni di GPFS sono certificate per alcuni kernel XEN RedHat, rendendo di fatto compatibile GPFS con la virtualizzazione XEN. Un test di compilazione e' stato fatto su una macchina SLC5.3 con kernel 2.6.18-128.1.1.el5xen.

Load balancing services

In testa ai server del cluster sono stati posti **2 switch Load Balancer Barracuda 340** in configurazione ridondata. Ciascuno switch e' dotato di 2 interfacce Gigabit, una connessa verso la WAN e l'altra connessa verso i server.

I Barracuda offrono i servizi dei “**real server**” attraverso un **virtual server** con indirizzo pubblico e possono distribuire il carico dei servizi attraverso politiche differenti (Round Robin,Weighted Least Connections, carico delle CPU letto via SNMP, numero di sessioni, ecc). Gli switch sono essenzialmente macchine Linux dai costi contenuti su cui e' installato un software basato su LVS (Linux Virtual Server), facilmente configurabile via web.

La connessione tra gli switch barracuda ed i real server e' in modalita' **Route Path**, con qualche modifica rispetto alla configurazione “standard” , dove e' previsto che i real server siano collegati allo switch in classe privata e che abbiano come default l'indirizzo IP dell'interfaccia del Barracuda. Nel nostro caso una opportuna combinazione di **iproute2 + iptables** marca tutti e soli i pacchetti destinati al traffico verso lo switch e attribuisce loro l'indirizzo IP dello switch come default. In questa maniera e' possibile configurare una seconda scheda di rete con IP e gateway pubblici registrati sulla nostra LAN, permettendo cosi' agli utenti di connettersi direttamente ai real server del cluster in modo analogo a quanto avviene con il cluster di lxplus.cern.ch.



High Availability

Nella progettazione del cluster si e' cercato di evitare i possibili “single point of failure”, vediamo quali:

•File system.
I server del cluster condividono due file-system GPFS (**/system** e **/user**) replicati (data e metadata) su due sistemi **Xyratex 5412** con 12 dischi Hitachi SATA II da 1 TB. Ciascun filesystem e' quindi composto da 2 Network Shared Disk (**NSD)** creati a partire da array di dimensione analoga sui due storage (**RAID6** per /system e **RAID5** per /user); tutti e 4 gli **NSD** appartengono a “failure group” diversi. In questa maniera il filesystem e' replicato su due storage diversi ed e' in grado di operare anche in caso totale “failure” di uno dei due.

• Ridondanza dei link Fibre Channel
Le connessioni degli storage e dei server sono opportunemante bilanciate e distribuite sui due switch FC.
La ridondanza delle connessione FC si traduce nella molteplicita' dei PATH (e delle device) con cui i singoli array degli storage possono essere visti dai server. Il software **device-mapper-multipath** gestisce questa ridondanza attraverso una opportuna politica di failover.

•Hardware.
I server (biprocessori Xeon e Opteron) sono dotati di alimentazione ridondata e dischi di sistema in **RAID1**. Tutti i sistemi di storage (due **Xyratex 5412** e un **InforTrend A16F-1211**) sono in configurazione ridondata con doppio controller e doppia alimentazione. I link FC sono distribuiti su due switch fibre channel **QLOGIC 5600** a 16 porte da 4Gbps. I due switch Load Balancer **Barracuda 340** sono a loro volta in configurazione “**High Availability**”.

•Backup
Il terzo sistema di storage (**A16F-1211-R**), con 16 dischi Hitachi da 1TB SATA I costituisce l'unita' backup in linea del cluster. Anche in questo caso l'unita' e' connessa in multipath ai server e lo spazio su disco e' diviso in due array (**RAID5**) su cui sono stati creati due filesystem GPFS (**/bsystem** e **/buser)**. L'idea e' quella di eseguire un backup incrementale ogni settimana ed eventualmente salvare gli snapshot periodici dei filesystem GPFS. Completa il sistema di backup una **Tape Library LTO Tandberg StorageLibrary T40** (9.6TB nativi, 19.2TB in compressione, 24 slot espandibili a 40, 1 drive LTO-4 FC nativo) connessa via Fibre Channel al cluster.