LBnamed: una soluzione per il DNS load-balancing

Gennaro Tortone [tortone@na.infn.it]

Workshop CCR INFN – La Biodola – Maggio 2002

Introduzione

Sempre più spesso viene adottata la soluzione del cluster per offrire servizi agli utenti (mail, calcolo, public login);

Tuttavia, sia per l'alto numero di utenti che per le esigenze di questi ultimi (job CPU intensive, job I/O intensive), si avverte l'esigenza di fornire un accesso bilanciato a ciascuna delle macchine che compongono il cluster. Bilanciare l'accesso alle singole macchine prevede la ricerca di un parametro di confronto, che definiamo carico (utilizzo di CPU, disponibilità spazio disco, traffico di rete), secondo il quale deve essere realizzato il bilanciamento;

Una possibile soluzione e' quella di realizzare il "dispatching" a livello del DNS;

Soluzione n. 1

BIND inside

Nelle recenti versioni di BIND è possibile definire un alias che punta a diversi host e, secondo una politica di tipo Round Robin, vengono restituiti ciclicamente gli indirizzi IP dell'insieme di server definiti.

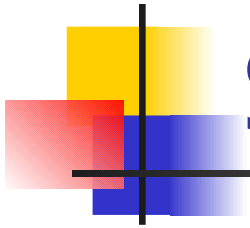
```
host1    IN A    192.168.1.1
host2    IN A    192.168.1.2
host3    IN A    192.168.1.3
host4    IN A    192.168.1.4
loadbal  IN CNAME host1 host2 host3 host4
```

PRO

- soluzione semplice e veloce;

CONTRO

- la politica di restituzione dell'indirizzo IP è puramente ciclica e se un host dell'insieme `loadbal` non è raggiungibile allora falliranno il 25% degli accessi verso `loadbal.dominio.it`.
- non tiene conto del fattore di “load” e del “fail over protection”



Soluzione n. 2

BIND patching

Esistono delle patches da applicare ai sorgenti del BIND (**dlbDNS**) che definiscono un nuovo record di risorsa da utilizzare per la definizione dell'insieme degli host; su ogni host è necessario avviare un deamon che risponde in modo asincrono al polling del nameserver primario fornendo a quest'ultimo i fattori di carico rilevati (utilizzo di CPU, swap space, numero di utenti collegati, numero di processi, etc.).

PRO

- load balancing;
- fail over protection;

CONTRO

- patch disponibile solo per alcune versioni di BIND;
- ogni volta che viene aggiornato il BIND sul nameserver primario e secondario bisogna adattare la patch (!) e applicarla ai sorgenti;

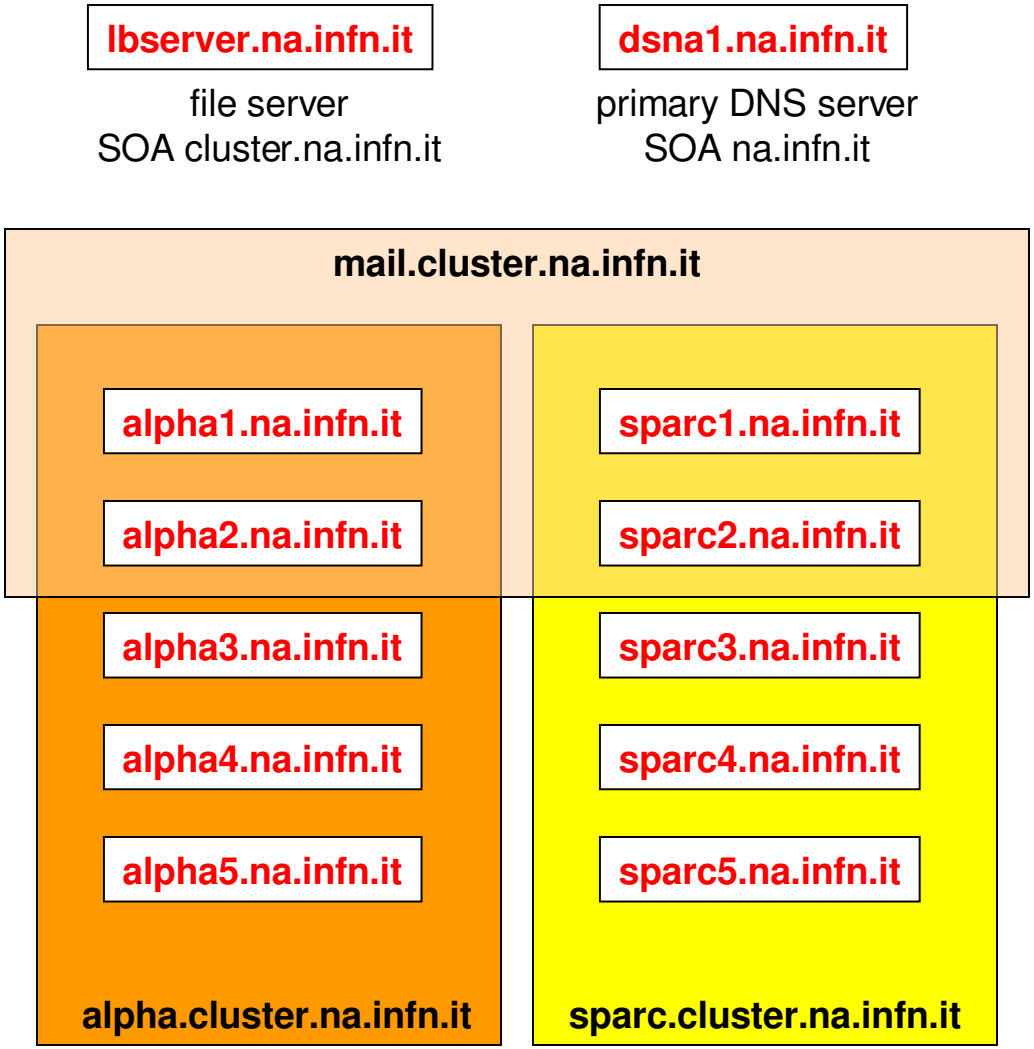
LBnamed

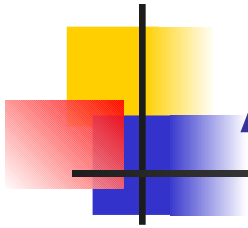
Introduzione

LBnamed è un set di programmi in C e Perl scritti da **Roland J. Schemers** (Stanford University) e rappresenta una soluzione al problema del "load balancing" in un cluster;
LBnamed è un vero e proprio nameserver che consente di creare gruppi dinamici di hosts dove un host può anche comparire in diversi gruppi;

Scenario tipico (cluster.na.infn.it = LB zone)

ALIAS DEL CLUSTER	SET DI SERVER
alpha.cluster.na.infn.it	alpha1, alpha2, alpha3, alpha4, alpha5
sparc.cluster.na.infn.it	sparc1, sparc2, sparc3, sparc4, sparc5
mail.cluster.na.infn.it	alpha1, alpha2, sparc1, sparc2


lbserver.na.infn.it
file server
SOA cluster.na.infn.it
dsna1.na.infn.it
primary DNS server
SOA na.infn.it
mail.cluster.na.infn.it
alpha1.na.infn.it
alpha2.na.infn.it
alpha3.na.infn.it
alpha4.na.infn.it
alpha5.na.infn.it
alpha.cluster.na.infn.it
sparc1.na.infn.it
sparc2.na.infn.it
sparc3.na.infn.it
sparc4.na.infn.it
sparc5.na.infn.it
sparc.cluster.na.infn.it



Architettura

Il server LBnamed

Il server LBnamed consiste di due programmi: lbnamed e poller;

Poller

Il daemon poller, in esecuzione sul nameserver lbnamed, contatta, ad intervallo di tempo specificato, il client daemon in esecuzione su ogni host dell'insieme cluster.

Gli host da contattare vengono specificati in un file di configurazione e, periodicamente, il poller manda una richiesta e riceve una risposta in modo asincrono.

Se il poller non riceve informazioni da un determinato host, rimuove quest'ultimo dalla lista degli host disponibili.

Al termine della ricezione delle risposte da parte delle macchine che compongono il cluster, esegue il dump in un file di tutte le informazioni raccolte e manda un segnale al daemon lbnamed che, in questo modo, viene aggiornato sulla situazione attuale del cluster.

Architettura

Il poller daemon è responsabile del calcolo del “carico" di ciascun host dell'insieme cluster. La formula usata per determinare il carico di un host è la seguente:

`$tot_user:` numero totale di utenti collegati;
`$uniq_user:` numero totale di utenti “unici” collegati;
`$load:` upime dell'ultimo minuto moltiplicato per 100;
`$WT_PER_USER:` peso attribuito a ciascun utente;
`$USER_PER_LOAD_UNIT:` fattore per cui moltiplicare $load;
`$fudge:` peso aggiuntivo per gli utenti collegati più volte allo stesso host;
`$weight:` carico dell'host;

```
$WT_PER_USER = 100;
$USER_PER_LOAD_UNIT = 3;
$fudge = ($tot_user - $uniq_user) * ($WT_PER_USER/5);
$weight = $uniq_user * $WT_PER_USER + ($USER_PER_LOAD_UNIT * $load)
          + $fudge;
```

La formula favorisce gli host con basso carico e basso numero di "login unici"

Architettura

Esempio di file generato dal poller:

```
1322 alpha1 192.168.1.1 alpha mail
1244 alpha2 192.168.1.2 alpha mail
 893 sparc1 192.168.1.3 sparc
1002 sparc2 192.168.1.4 sparc mail
...
```

LBnamed

LBnamed è un nameserver che non accetta query ricorsive; risponde solo a richieste di indirizzi IP per gli alias che sono definiti nella sua authority. A seguito di una richiesta sulla porta 53, risponde al mittente con un pacchetto DNS standard in cui il TTL per il record di risorsa (CNAME) fornito è settato ad un valore basso (es. ≤ 5). Questo per evitare il caching, da parte del client, dell'indirizzo IP richiesto.
Lo script LBnamed legge il file di configurazione generato dal poller e carica il suo contenuto in differenti strutture dati. Quando arriva una richiesta per un determinato gruppo viene restituito l'host con il peso minore e viene incrementato di ($WT_PER_USER * 2) il rispettivo peso. In questo modo viene aggiornato il peso dell'host anche durante il periodo di raccolta dati del poller;

Architettura

Di seguito viene riportato il risultato di una prima query verso il cluster alpha.cluster.na.infn.it:

```
# dig alpha.cluster.na.infn.it
; <<>> DiG 2.1 <<>> alpha.cluster.na.infn.it
    ;; res options: init recurs defnam dnsrch
    ;; got answer:
    ;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 6
    ;; flags: qr aa rd ra; Ques: 1, Ans: 2, Auth: 0, Addit: 0
    ;; QUESTIONS:
    ;; alpha.cluster.na.infn.it, type = A, class = IN
;; ANSWERS:
    alpha.cluster.na.infn.it. 2 CNAME alpha1.na.infn.it.
    alpha1.na.infn.it. 3600 A 192.84.134.83
;; Total query time: 60 msec
    ;; FROM: 192.84.134.50 to SERVER: default -- 192.135.13.7
    ;; WHEN: Sat Sep 30 11:13:57 2000
    ;; MSG SIZE sent: 42 rcvd: 114
```

Ecco il risultato della seconda query

```
# dig alpha.cluster.na.infn.it
;; ANSWERS:
    alpha.cluster.na.infn.it. 2 CNAME alpha3.na.infn.it.
    alpha3.na.infn.it. 3600 A 192.84.134.84
```

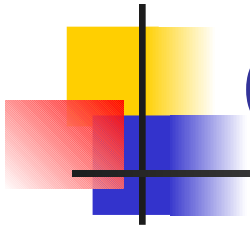
Architettura

LBCD (Load Balancing Client Daemon)

LBCD è un daemon scritto in C che deve essere eseguito dagli host appartenenti al cluster al fine di rispondere alle richieste del poller. Il protocollo usato è molto semplice e viene implementato tramite UDP;

Ecco la definizione del pacchetto:

```
#define PROTO_PORTNUM 4330
typedef struct {
        u_short version;        /* protocol version */
        u_short id;             /* requestor's uniq request id */
        u_short op;             /* operation requested */
        u_short status;         /* set on reply */
    } P_HEADER,*P_HEADER_PTR;
typedef struct {
        P_HEADER h;
        u_int boot_time;
        u_int current_time;
        u_int user_mtime;       /* time user information last changed */
        u_short l1;             /* (int) (load*100) */
        u_short l5;
        u_short l15;
        u_short tot_users;      /* total number of users logged in */
        u_short uniq_users;     /* total number of uniq users */
        u_char on_console;      /* true if someone on console */
        u_char reserved;        /* future use, padding... */
    } P_LB_RESPONSE, *P_LB_RESPONSE_PTR;
```



Configurazione

Passo 1

Per il load balancing nameserver deve essere dichiarato un sottodominio all'interno del nameserver primario.

Questo può essere fatto semplicemente con il record di risorsa (NS).

```
cluster     IN NS     lbserver
```

In questo modo deleghiamo le risoluzioni per il dominio `cluster.na.infn.it` a `lbserver.na.infn.it` su cui sarà installato LBnamed. Quando al nameserver primario giungeranno richieste del tipo `alpha.cluster.na.infn.it` esso le ruoterà a `lbserver.na.infn.it`.

Configurazione

Passo 2

Compilare LBCD, copiarlo su ciascun nodo della zona bilanciata ed avviarlo;

Passo 3

Configurare il daemon LBnamed tramite il file lbnamed.conf;

Passo 4

Configurare il poller tramite il file cluster.config;

Passo 5

Avviare il daemon Lbnamed e il poller sul server della zona bilanciata (`lbserver.na.infn.it`);

lbnamed.conf

```
BEGIN {
        $poller_sleep = 120;
        $poller_config = “cluster.config";
        $hostmaster = “gennaro.tortone.na.infn.it";
    }

    LBDB::add_static(“cluster.na.infn.it",T_SOA,rr_SOA(hostname,
    $hostmaster,time,86400,86400,86400,0));

    LBDB::add_dynamic(“cluster.na.infn.it" => \&handle_best_request);

    # dynamic domain handlers...

    sub handle_best_request { [...] }

[...]
```

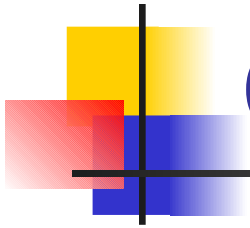
cluster.config

Le linee del file di configurazione del poller (`cluster.config` nel nostro caso) devono essere del tipo:

```
host divisore-peso gruppo1 [gruppo2 ...]
```

Il campo `divisore-peso` rappresenta un fattore per cui viene diviso il peso di un determinato host. Tale fattore può essere utile al fine di determinare delle priorità diverse all'interno degli host di un gruppo. Ad un host con alte performance possiamo assegnare un divisore uguale a 2 in modo da dimezzare il suo peso calcolato dal poller.

```
alpha1    2    alpha mail
alpha2    1    alpha mail
alpha3    2    alpha
alpha4    1    alpha
alpha5    1    alpha
sparc1    1    sparc mail
sparc2    3    sparc mail
sparc3    1    sparc
sparc4    2    sparc
sparc5    1    sparc
```



Conclusioni

Vantaggi dell'implementazione del "load balancing" con LBnamed

- nessun cambiamento ai sorgenti del BIND;
- facilità di configurazione;
- load balancing & fail over protection;
- possibilità di creare sottodomini gestiti da routines personalizzate;
- portabilita' del codice (Perl5 + Ansi C);

Riferimenti

Home page

http://www.stanford.edu/~riepel/lbnamed

Documentazione (in italiano)

http://people.na.infn.it/~tortone/lbnamed-docs