Linux Cluster HOWTO

Ram Samudrala (me@ram.org) v1.5, 5 settembre 2005

Come configurare cluster Linux ad alte prestazioni

Contents

1 Introduzione

Questo documento descrive come abbiamo configurato i cluster ad alte prestazioni che utilizziamo per la nostra ricerca .

L'utilizzo dell'informazione seguente è a proprio rischio. Declino ogni responsabilità per qualsiasi cosa si faccia dopo aver letto questo HOWTO. L'ultima versione di questo HOWTO sarà sempre disponibile all'indirizzo http://www.ram.org/computing/linux/linux_cluster.html .

A differenza di altra documentazione che spiega come configurare cluster in generale, questa è una descrizione specifica di come è configurato il nostro laboratorio e include non solo dettagli sugli aspetti computazionali, ma anche su desktop, portatili e server pubblici. Ciò è fatto principalmente per uso locale, ma l'ho pubblicato sul web perché ho ricevuto molti messaggi di posta elettronica che chiedevano le stesse informazioni. Anche oggi, pianificando un altro cluster a 64 nodi, rilevo scarsità di informazione su come assemblare esattamente i componenti per formare un nodo che funzioni in modo affidabile sotto Linux, vale a dire informazione non solo sui nodi di calcolo, ma anche sull'hardware che deve funzionare bene con i nodi affinché sia possibile una ricerca produttiva. La principale utilità di questo HOWTO è che riporta quale tipo di hardware funziona bene con Linux e quale no.

Traduzione a cura di Amelia de Vivo. Revisione di Giulio Daprelà (giulio at pluto.it)

2 Hardware

Questo paragrafo copre le scelte hardware che ho fatto. A meno che non sia riportato nel paragrafo 4.4 (problemi hardware noti), si assuma che tutto funzioni *realmente* bene.

L'installazione dell'hardware è anche abbastanza facile, a meno che non sia esplicitamente riportato, e la maggior parte dei dettagli è coperta dai manuali. In tutti i paragrafi l'hardware è elencato in ordine di acquisto (il più recente è elencato per primo).

2.1 I nodi

32 macchine hanno la seguente configurazione:

- 2 CPU XEON 2.66GHZ 533FSB

- Scheda madre Supermicro 6013A-T, 1U
- 2 moduli RAM 512MB PC2100 DDR REG ECC
- 1 HD 80 GB SEA 7200 RPM SATA
- 1 HD 250 GB SEA 7200 RPM SATA

32 macchine hanno la seguente configurazione:

- 2 CPU XEON 2.4GHZ 533FSB
- Scheda madre Supermicro X5DPR-1G2
- 2 moduli RAM 512 MB PC2100 DDR REG ECC
- 1 HD 40GB SEA 7200 RPM
- 1 HD 120GB SEA 7200 RPM
- CDROM Supermicro Slim 24X
- Alloggiamento CSE-812 400 C/B 1U

32 macchine hanno la seguente configurazione:

- 2 CPU AMD Palamino MP XP 2000+ 1.67 GHz
- Scheda madre Asus A7M266-D w/LAN Dual DDR
- 2 moduli RAM Kingston 512MB PC2100 DDR-266MHz REG ECC
- 1 HD 41 GB Maxtor 7200 RPM ATA100
- 1 HD 120 GB Maxtor 5400 RPM ATA100
- CDROM Asus CD-A520 52x
- Floppy drive 1.44 MB
- ATI Expert 2000 Rage 128 32 MB
- Alloggiamento mezza torre IN-WIN P4 300ATX
- Alimentatore Enermax P4-430ATX

32 macchine hanno la seguente configurazione:

- 2 CPU AMD Palamino MP XP 1800+ 1.53 GHz
- Scheda madre Tyan S2460 Dual Socket-A/MP
- RAM Kingston 512 MB PC2100 DDR-266MHz REG ECC

- 1 HD 20 GB Maxtor UDMA/100 7200 RPM
- 1 HD 120 GB Maxtor 5400 RPM ATA100
- CDROM Asus CD-A520 52x
- Floppy drive 1.44 MB
- Scheda video ATI Expert 98 8 MB AGP
- Alloggiamento mezza torre IN-WIN P4 300ATX
- Scheda di rete Intel PCI PRO-100 10/100Mbps
- Alimentatore Enermax P4-430ATX

32 macchine hanno la seguente configurazione:

- 2 CPU Pentium III 1 GHz Intel
- Scheda madre Supermicro 370 DLE Dual PIII-FCPGA
- 2 moduli RAM 256 MB 168-pin PC133 Registered ECC Micron
- 1 HD 20 GB Maxtor ATA/66 5400 RPM
- 1 HD 40 GB Maxtor UDMA/100 7200 RPM
- CDROM Asus CD-S500 50x
- Floppy drive 1.44 MB
- Scheda video ATI Expert 98 8 MB PCI
- Alloggiamento mezza torre IN-WIN P4 300ATX

2.2 I server

Due server per uso esterno (diffusione di informazioni) con le seguenti configurazioni:

- 2 CPU AMD Opteron 240 1.4 GHz
- Scheda madre RIOWORKS HDAMB DUAL OPTERON
- 4 moduli RAM KINGSTON 512 MB PC3200 REG ECC
- HD 80 GB MAX 7200 RPM UDMA 133
- 6 HD 200 GB WD 7200 RPM 8 MB
- CDROM ASUS 52X CD-A520
- Floppy drive 1.44 MB

- Alloggiamento Antec 4U22ATX550EPS 4U

- 2 CPU AMD Palamino MP XP 2000+ 1.67 GHz
- Scheda madre Asus A7M266-D w/LAN Dual DDR
- 4 moduli RAM Kingston 512 MB PC2100 DDR-266MHz REG ECC
- CDROM Asus CD-A520 52x
- 1 HD 41 GB Maxtor 7200 RPM ATA100
- 6 HD 120 GB Maxtor 5400 RPM ATA100
- Floppy drive 1.44 MB
- ATI Expert 2000 Rage 128 32 MB
- Alloggiamento mezza torre IN-WIN P4 300ATX
- Alimentatore Enermax P4-430ATX

2.3 I Desktop e i Terminali

Abbiamo identificato almeno due tipi di utenti per i nostri cluster: quelli che hanno bisogno (ovvero traggono vantaggio da) di potenza computazionale permanente e locale e spazio su disco in congiunzione con il cluster per velocizzare il calcolo, e quelli che hanno solo bisogno della potenza computazionale del cluster. Ai primi sono assegnati dei "desktop" che essenzialmente sono macchine ad alte prestazioni, e ai secondi sono assegnati dei "terminali" stupidi. I nostri desktop son in genere macchine a due o quattro processori che montano CPU Opteron a 1.6 GHz, attualmente di alto profilo, hanno fino a 10 GB di RAM e oltre 1 TB di spazio di disco locale. I nostri terminali sono essenzialmente macchine da cui un utente può collegarsi e lanciare job sui cluster. In questa configurazione le persone possono anche usare dei portatili come terminali stupidi.

2.4 Varie/accessori

Generalmente usiamo/preferiamo monitor Viewsonic, mouse Microsoft Intellimouse e tastiere Microsoft Natural. Questi in genere hanno funzionato in modo abbastanza affidabile per noi.

2.5 Hardware per mettere tutto insieme

Per l'accesso visuale ai nodi, inizialmente usavamo switch KVM con un monitor economico per connettersi e "guardare" tutte le macchine. Sebbene fosse una bella soluzione, non scalava. Attualmente portiamo un piccolo monitor in giro e attacchiamo dei cavi quando serve. Quello che ci serve è un piccolo monitor portatile che si può collegare dietro a un PC (alimentato con una stilo, come un Palmare).

Per la rete, in genere usiamo switch Netgear e Cisco.

2.6 Costi

Abbiamo acquistato il nostro hardware presso Hard Drives Northwest (http://www.hdnw.com). Per ciascun nodo del cluster (contenente due processori), abbiamo speso circa 1500-2000 dollari, tasse incluse. Generalmente, il nostro scopo è tenere il costo di ciascun processore sotto i 1000 dollari (compreso l'alloggiamento).

3 Software

3.1 Sistema operativo: Linux, naturalmente!

I kernel e le distribuzioni in uso sono i seguenti:

- Kernel 2.2.16-22, distribuzione KRUD 7.0
- Kernel 2.4.9-7, distribuzione KRUD 7.2
- Kernel 2.4.18-10, distribuzione KRUD 7.3
- Kernel 2.4.20-13.9, distribuzione KRUD 9.0
- Kernel 2.4.22-1.2188, distribuzione KRUD 2004-05

Queste distribuzioni vanno molto bene per noi perché gli aggiornamenti ci vengono mandati su CD e non c'è nessun bisogno di connessioni esterne. Inoltre sembrano "più pulite" delle normali distribuzioni Red Hat e la configurazione è estremamente stabile.

3.2 Software di rete

Noi usiamo Shorewall 1.3.14a (http://www.shorewall.net) per il firewall.

3.3 Software per il calcolo parallelo

Noi usiamo software proprietario per parallelizzare le applicazioni, ma abbiamo fatto esperimenti con PVM e MPI . Secondo me l'overhead di queste librerie preconfezionate è troppo alto. Raccomando di scrivere codice specifico per l'applicazione che vi interessa (è un opinione personale).

3.4 Costi

Linux e la maggior parte del software che gira sotto Linux sono liberamente utilizzabili.

4 Installazione, configurazione e manutenzione

4.1 Configurazione dei dischi

Questo paragrafo descrive le strategie di partizionamento dei dischi. Il nostro scopo è mantenere le strutture virtuali delle macchine organizzate in modo tale che siano tutte logiche. Stiamo constatando che i mapping fisici alle strutture logiche non sono sostenibili quando l'hardware e il software (sistema operativo) cambiano. Attualmente, la nostra strategia è la seguente:

```
nodi dei cluster:

partizione 1 sul disco di sistema              - swap   (2 * RAM)
partizione 2 sul disco di sistema              - /      (il resto dello spazio del disco)
partizione 1 su un disco addizionale    - /maxa  (tutto il disco)

server:

partizione 1 sul disco di sistema              - swap  (2 * RAM)
partizione 2 sul disco di sistema              - /     (4-8 GB)
partizione 3 sul disco di sistema              - /home (il resto dello spazio del disco)
partizione 1 sul disco addizionale 1    - /maxa (tutto il disco)
partizione 1 sul disco addizionale 2    - /maxb (tutto il disco)
partizione 1 sul disco addizionale 3    - /maxc (tutto il disco)
partizione 1 sul disco addizionale 4    - /maxd (tutto il disco)
partizione 1 sul disco addizionale 5    - /maxe (tutto il disco)
partizione 1 sul disco addizionale 6    - /maxf (tutto il disco)
partizione 1 su disco/i addizionale/i   - /maxg (spazio totale sui dischi)

desktop:

partizione 1 sul disco di sistema              - swap   (2 * RAM)
partizione 2 sul disco di sistema              - /      (4-8 GB)
partizione 3 sul disco di sistema              - /spare (il resto dello spazio del disco)
partizione 1 sul disco addizionale 1    - /maxa (tutto il disco)
partizione 1 su disco/i addizionale/i   - /maxb (spazio totale sui dischi)
```

Si noti che nel caso di server e desktop, maxg e maxb possono essere un singolo disco o un agglomerato di dischi.

4.2 Configurazione dei pachetti software

Installare un insieme minimale di pacchetti per la collettività. Gli utenti possono configurare i desktop come vogliono, fermo restando che la struttura virtuale è mantenuta come descritto sopra.

4.3 Installazione e manutenzione del sistema operativo

4.3.1 Strategia personale di clonazione

Io credo nell'utilità di avere un sistema completamente ditribuito. Ciò significa che ogni macchina ha una copia del sistema operativo. Installare il SO su ciascuna macchina manualmente è scomodo. Per ottimizzare questo processo, prima installo e configuro una macchina esattamente come voglio. Poi creo un file tar compresso dell'intero sistema e lo metto su un CD-ROM di avviamento che infine clono su ogni macchina del mio cluster.

I comandi che uso per creare il tar sono i seguenti:

```
tar -czvlps --same-owner --atime-preserve -f /maxa/slash.tgz /
```

Uso uno script che si chiama `go` che prende in input un numero di macchina, scompatta il file `slash.tgz` dal CD-ROM e sostituisce l'hostname e l'indirizzo IP nelle locazioni appropriate. Una versione dello script `go` e i relativi file input si trovano a: http://www.ram.org/computing/linux/linux/cluster/ . Questo script dovrà essere modificato in base alla configurazione del vostro cluster.

Per fare questo lavoro uso Custom Rescue Disk di Martin Purschke (http://www.phenix.bnl.gov/~purschke/RescueCD/) per creare su un CD di avviamento un'immagine contenente il file .tgz che rappresenta il sistema clonato, come pure lo script `go` e altri file relativi. Tutto ciò è impresso su un CD-ROM.

Ci sono parecchi documenti che descrivono come creare un proprio CD di avviamento, compreso il Linux Bootdisk HOWTO (http://www.linuxdoc.org/HOWTO/Bootdisk-HOWTO/), che contiene anche dei link ad altri dischi di boot/root preconfezionati.

Ora avete un sistema grazie al quale tutto quello che dovete fare è inserire un CDROM, accendere la macchina, andare a prendere un caffè (o una lattina di coca) e tornare a vedere un clone completo. Ripetete questo processo per tutte le macchine che avete. Questa procedura ha funzionato meravigliosamente bene per me e se avete qualcun altro che effettivamente fa il lavoro (di inserire e rimuovere i CDROM), allora è l'ideale. Nel mio sistema, specifico l'indirizzo IP specificando il numero della macchina, ma questo potrebbe essere completamente automatizzato con l'uso di DHCP.

Rob Fantini (rob@fantinibakery.com) ha contribuito alle modifiche dei suddetti script che ha usato per clonare un sistema Mandrake 8.2 accessibili all'indirizzo http://www.ram.org/computing/linux/cluster/fantini_contribution.tgz .

4.3.2 Pacchetti per la clonazione e la manutenzione

FAI FAI (http://www.informatik.uni-koeln.de/fai/) è un sistema automatico per installare un sistema operativo GNU/Linux Debian su un cluster di PC. Potete prendere uno o più PC vergini, accenderli e dopo pochi minuti Linux sarà installato, configurato e funzionante sull'intero cluster, senza necessità di alcuna interazione.

SystemImager SystemImager (http://systemimager.org) è un software che automatizza l'installazione di Linux, la distribuzione del software e la rapida produzione.

4.3.3 DHCP e indirizzi IP fissi

Se avete configurato il DHCP, non avete bisogno di risistemare gli indirizzi IP e la relativa parte può essere rimossa dallo script `go`.

Il DHCP ha il vantaggio di non farvi affatto impazzire con gli indirizzi IP, posto che il server DHCP sia configurato in modo esatto. Ha però lo svantaggio di dipendere da un server centralizzato (e come ho detto, io tendo a distribuire le cose quanto più è possibile). Inoltre, legare gli indirizzi delle schede ethernet agli indirizzi IP può essere scomodo se volete sostituire macchine o cambiare hostname spesso.

4.4 Problemi hardware noti

L'hardware in generale ha funzionato veramente bene. Problemi specifici sono elencati di seguito:

Le macchine AMD bi-processori 1.2 GHz dissipano molto calore. Due in una stanza fanno aumentare significativamente la temperatura. Mentre potrebbero essere adatte come desktop, il raffreddamento e il consumo di energia quando le si usa come parte di un grosso cluster vanno considerati. La configurazione AMD Palmino precedentemente descritta sembra funzionare veramente bene, ma di sicuro raccomando di avere due ventole nell'alloggiamento–questo ha risolto tutti i nostri problemi di instabilità.

4.5 Problemi software noti

Alcuni eseguibili tar a quanto pare non creano un file tar nel modo che ci si aspetta (specialmente per quanto riguarda il riferimento e il de-riferimento dei link simbolici). La soluzione che ho trovato è usare un eseguibile tar che funzioni correttamente, come quello fornito da RedHat 7.0.

5 Eseguire task sul cluster

Questo paragrafo è ancora in evoluzione man mano che evolve l'utilizzo del cluster, ma finora tendiamo a scrivere insiemi proprietari di funzioni message passing per la comunicazione dei processi su macchine diverse.

Molte applicazioni, particolarmente nell'area della genomica computazionale, sono massivamente e banalmente parallelizzabili, vale a dire che si può ottenere una perfetta distribuzione ripartendo i task equamente tra le macchine (per esempio, quando si analizza un intero genoma usando una tecnica che opera su un singolo gene/proteina, ciascun processore in un dato istante può lavorare su un gene/proteina indipendente da tutti gli altri processori).

Finora non abbiamo trovato necessario usare un sistema professionale a code, ma ovviamente dipende fortemente dal tipo di applicazioni che si vogliono eseguire.

5.1 Benchmark grossolani

Per l'unico programma più importante che eseguiamo (il nostro programma di simulazione *ab initio* del ripiegamento delle proteine), utilizzando il processore Pentium 3 a 1 GHz come punto di riferimento, in media:

```
Xeon    1.7 GHz è circa il 22% più lento
Athlon  1.2 GHz è circa il 36% più veloce
Athlon  1.5 GHz è circa il 50% più veloce
Athlon  1.7 GHz è circa il 63% più veloce
Xeon    2.4 GHz è circa il 45% più veloce
Xeon    2.7 GHz è circa il 80% più veloce
Opteron 1.4 GHz è circa il 70% più veloce
Opteron 1.6 GHz è circa il 88% più veloce
```

Sì, l'Athlon 1.5 GHz è più veloce dello Xeon 1.7 GHz poiché lo Xeon esegue solo sei istruzioni per clock (IPC) mentre l'Athlon esegue nove IPC (fate voi i conti!). Comunque questo è un confronto assolutamente non rigoroso perché gli eseguibili sono stati entrambi compilati sulle macchine (quindi la qualità delle librerie matematiche per esempio avrà un suo impatto) e l'hardware di supporto è diverso.

5.2 Uptime

Queste macchine sono incredibilmente stabili sia in termini di hardware che di software una volta che sono state debuggate (in genere qualcuna in un nuovo gruppo di macchine ha problemi hardware), funzionando costantemente sotto carichi molto pesanti. Di seguito è dato un esempio. I reboot sono generalmente avvenuti quando è scattato un interruttore.

```
2:29pm  per 495 giorni,  1:04,  2 utenti,  carico medio: 4.85, 7.15, 7.72
```

6 Ringraziamenti

Le seguenti persone hanno utilmente contribuito a questo HOWTO:

- Michal Guerquin (Michal Guerquin)
- Michael Levitt (Michael Levitt)

7 Bibliografia

I seguenti documenti potrebbero essere utili—sono link a progetti che fanno uso di cluster ad alte prestazioni:

- Ram Samudrala's research page (che descrive il tipo di ricerca fatto con questi cluster)
- RAMP web page
- RAMBIN web page