Un Cluster ad alte prestazioni per il progetto AMS

Pavol Bobik (*), Matteo Boschini (**), Luca Carbone (*), Davide Grandi (*)
(*) *INFN, Milano*
(**) *CILEA, Segrate*

Abstract
Viene presentato lo sviluppo di un cluster ad architettura PC Linux per calcolo ad alte prestazioni nell'ambito del progetto AMS.

Keywords: Telematica, Linux, Cluster, Calcolo ad alte prestazioni.

Introduzione

Il progetto AMS, cui il CILEA partecipa in maniera attiva dal 1997 in collaborazione con la sezione di Milano dell'I.N.F.N. è descritto in dettaglio altrove [1]-[6]. Si intende qui dare una breve descrizione delle risorse di calcolo ad alte prestazioni necessarie alle analisi di fisica e come si sia fatto fronte ad esse implementando un cluster di PC.

L'analisi dei dati

Studiare la traiettoria di particelle di origine cosmica all'interno della regione di spazio influenzata dal campo magnetico terrestre, nota come magnetosfera è necessario per poter spiegare la natura dei raggi cosmici rivelati in prossimità della Terra, e riveste quindi un ruolo fondamentale per l'analisi dei dati di AMS.
Per uno studio analitico delle particelle rivelate da AMS (raggi cosmici) è necessario utilizzare un programma che riproduca la traiettoria delle particelle cariche nel complesso campo magnetico presente nella magnetosfera. Si considera l'effetto del campo principale (interno) con il modello IGRF 2000, e del campo solare (esterno), riprodotto dal nuovo modello Tsyganenko 96. L'evoluzione del campo esterno, trasportato dal vento solare, segue i cicli di attività del Sole, e modifica il campo magnetico totale che agisce sulle particelle, determinando la forma della magnetosfera terrestre. Con il programma di tracciamento si può calcolare la traiettoria e il periodo di permanenza in orbita delle particelle, riuscendo così a distinguere le primarie provenienti dai limiti della magnetosfera, da quelle secondarie provenienti dall'interazione dei raggi cosmici con gli strati alti dell'atmosfera.
Il programma di tracciamento viene utilizzato per calcolare la rigidità di taglio geomagnetico a cui sarà soggetto il rivelatore AMS una volta installato sulla Stazione Spaziale Interna-zionale ISSA. Infine, il medesimo programma di tracciamento viene utilizzato per discri-minare particelle primarie da particelle secondarie all'interno del campione dati raccolto da AMS durante la missione STS-91 del 1998.

Le esigenze di calcolo

Dalla breve descrizione data sopra, si può facilmente dedurre come tale analisi necessiti di alte prestazioni di calcolo, soprattutto se si tiene conto che si devono tracciare circa 200 particelle al secondo, con un tempo medio di tracciatura di 5 particelle al secondo[1], per un totale di 2.5×10^8 particelle. Tuttavia, le risorse necessarie sono essenzialmente di CPU, essendo l'I/O molto ridotto per questo tipo di applicazioni[2].
Si rende quindi necessaria la realizzazione di un sistema di calcolo ad alte prestazioni e a basso costo, di facile manutenzione e che usi tecnologie note e diffuse nell'ambito del calcolo astrofisico. Adottando una soluzione basata su

[1] Su un processore AMD-Athlon a 1.4 GHz.
[2] Si pensi, ad esempio, che 10 giorni di *Real Elapsed Time* di simulazione producono circa 1 KByte di output.

soluzioni *COF* (*Components Off the Shelf*) si risolve la problematica dei costi, mentre l'uso di Linux come sistema operativo garantisce la robustezza dell'intero sistema, nonché la sua scalabilità e interoperabilità con altri sistemi di calcolo usati da AMS.
Si sono tuttavia scartate implementazioni di clustering "pre-confezionate" o basate su kernel modificati, quali ad esempio OpenMosix[7], per mantenere il più possibile slegata la realizzazione del sistema da dettagli di sistema operativo.
Si è dunque implementato un cluster "beowulf old style"[8], con un nodo che funge da *master* e 20 nodi che fungono da *client*. Inoltre, poichè l'applicativo non richiede particolari risorse in termini di spazio disco, i clienti sono *disk-less*, e caricano il sistema operativo per mezzo del protocollo EtherBoot[9].
La configurazione è quindi la seguente:

Master

- Kernel 2.4.9-12
- RedHat 7.1
- 100 Gbyte area scratch esportata NFS
- DHCP server
- EtherBoot server.

Client

- No disk – no operating system
- Bootp request al Master
- O.S. montato via NFS al boot, tramite protocollo EtherBoot
- Area scratch montata via NFS.

Il cluster così costituito è in funzione dal Gennaio del 2002, inizialmente in modalità *floppy-boot*, e da Agosto 5 nodi sono in modalità *eprom-boot*[3].
Ad oggi il sistema ha avuto solo una *failure* sul master (guasto all'alimentatore) e due *failure* sui clienti (guasto alla RAM), con un *up-time* complessivo di 320 giorni.
In totale sono state tracciate 2.0 x 10^8 particelle da simulazione e 0.3 x 10^8 particelle del campione STS-91.

Gestione delle risorse

Trattandosi di un cluster disk-less, non è stato possibile implementare gli usuali sistemi di gestione di processi batch e di code, quali Condor[10] o OpenPBS[11]; si è quindi implementato un sistema di code/batch scritto in Perl5 e basato su un meccanismo *fifo* e un algoritmo di *round-robin*.
Essenzialmente, quando un job viene sottomesso al sistema, se questo trova una macchina disponibile e non trova job già in attesa di essere sottomessi, toglie dall'elenco delle macchine disponibili il nodo prescelto e sottomette il job; se non ci sono macchine disponibili, i job vengono messi in coda, infine, se ci sono job già in coda, questi vengono sottomessi per primi.
L'algoritmo di *scheduling* è tutt'ora basato solo su un carico minimo e un carico massimo dei nodi, ma attualmente è in fase di sviluppo un meccanismo più complesso che permette l'interoperabilità con un sistema OpenPBS con il quale sono governati i nodi di un cluster con dischi dedicato all'analisi MonteCarlo del gruppo.

Bibliografia

[1] M. Boschini, L. Trombetta *"Un DataBase Oracle per l'esperimento AMS"*, Bollettino del CILEA, n.64 Settembre 1998, pagg.22-25
[2] M. Boschini , L. Trombetta *"Il DataBase per AMS a un anno dal volo dello Space Shuttle STS-91"*, Bollettino del CILEA, n. 68 Giugno 1999, pag. 21
[3] URL: http://ams.cern.ch
[4] M. Boschini e aa.vv. *"An Oracle (c) database for the AMS experiment"*, NUCL PHYS B-PROC SUP 78, pagg. 727-731, August 1999
[5] M. Boschini *"Stato del Progetto AMS"*, Bollettino del CILEA, n. 78 Giugno 2001, pagg. 4-7
[6] M. Boschini, A. Favalli *"Stato del progetto AMS"*, Bollettino del CILEA, n. 84 Ottobre 2002, pagg. 53-56
[7] www.openmosix.org
[8] www.beowulf.org
[9] http://etherboot.sourceforge.net
[10] www.cs.wisc.edu/condor/
[11] www.openpbs.org

[3] Floppy-boot: le informazioni relative al Master sono contenute su floppy. Eprom-boot: le informazioni sono contenute su di una eprom appositamente programmata ed installata sulla scheda di rete (3C-905-C).