Michelangelo, un sistema ad alte prestazioni per la bioinformatica italiana

Claudio Arlandini
CILEA, Segrate

Abstract
Nell'ambito del Progetto LITBIO il CILEA ha installato un nuovo supercalcolatore dedicato esclusivamente ad applicazioni bioinformatiche, denominato *Michelangelo*. Si tratta di un cluster realizzato mediante tecnologie innovative, composto di 280 *core* Opteron e di una rete di interconnessione ad alte prestazioni Infiniband 4X. Il sistema ha una potenza di picco di 1,2 TFlop/s. Combinato con le altre risorse di calcolo ad alte prestazioni del CILEA ha permesso il ritorno del centro nella prestigiosa classifica TOP500.

In the framework of Project LITBIO CILEA installed a new supercomputer, named *michelangelo*, dedicated to bioinformatic applications. It is a cluster, realized with innovative solutions, composed of 280 Opteron cores and a high performance interconnection network Infiniband 4X. The system has a peak performance of 1.2 TFlop/s. Once combined with the other CILEA high performance computing resources it allowed the return of the Center in the TOP500 list.

Keywords: Supercomputing, Bioinformatica, Cluster.

Introduzione

Il CILEA è partner organizzativo del Laboratorio Interdisciplinare di Tecnologie Bioinformatiche (LITBIO) (fig. 1), un progetto di livello internazionale [1] finanziato dal Ministero Università e Ricerca, mediante un bando FIRB 2003 per il periodo 2005-2010, coordinato dal dott. Luciano Milanesi dell'Istituto di Tecnologie Biomediche del CNR (CNR-ITB), con lo scopo di incoraggiare la ricerca Genomica e Proteomica.

Fig. 1 – Il logo del Progetto LITBIO

Lo scopo generale della Bioinformatica può essere riassunto nella generazione di nuove conoscenze di tipo indicativo mediante il confronto tra dati genomici, proteomici e di trascrizione con modelli di più larga scala, che consentono nuove scoperte nelle basi molecolari della vita. I dati generati da vari progetti di sequenziamento del genoma formano la base per estrarre nuove classi di informazioni, per esempio sulla storia evolutiva delle famiglie geniche, su moduli genomici e network regolativi, e su una classificazione genotipica, invece che fenotipica, delle specie.

Il numero dei super-calcolatori dedicati all'analisi del genoma umano è in continuo aumento in tutto il mondo. Le necessità che hanno giustificato questo sforzo derivano dalla natura stessa del lavoro di compilazione, assemblaggio e comparazione di dati spesso frammentari, e dall'aumento della quantità di dati, che provoca una crescita esponenziale dei tempi di analisi.

Algoritmi più efficienti ed efficaci, analisi più globali e approfondite inevitabilmente cambieranno il modo in cui gli scienziati affrontano la

ricerca biomedica. Inoltre, la complessità delle informazioni, l'enorme quantità di dati, il continuo aggiornamento e la difficoltà nell'uso di diversi programmi di analisi rendono difficile il lavoro quotidiano degli utilizzatori, per i quali i sistemi e gli strumenti di oggi sono peraltro indispensabili per la ricerca, la modellizzazione, la produzione e la formazione.

Viene quindi chiaramente percepito il bisogno di creare una piattaforma computazionale solida, completa e multidisciplinare, di disegnare e saper applicare software avanzato per *data-mining* per l'analisi di grandi collezioni di dati, di preparare una classe di professionisti orientati alla ricerca, con competenze specifiche, dedicati alla soluzione dei problemi e a stretto contatto con quei laboratori di ricerca che producono dati.

Per rispondere a queste esigenze è stato recentemente installato presso il CILEA un adeguato sistema di calcolo denominato *Michelangelo*, la cui struttura e dimensione evolverà nel corso del progetto.

Scopo di questo articolo è fornire alcuni dettagli tecnici sulla struttura e configurazione del cluster.

La configurazione hardware

Michelangelo è un sistema con architettura a cluster, fornito dal partner tecnologico del progetto, Exadron [2], del gruppo Eurotech SpA.

Al costruttore è stata richiesta una piattaforma a 64 bit con uno spazio disco di dimensioni sufficienti a gestire le basi di dati prodotte dai laboratori del progetto e con elevate prestazioni in lettura/scrittura, una interconnessione di rete ad alte prestazioni capace di gestire efficientemente simulazioni parallele ad alto numero di processori. È necessaria inoltre un'elevata flessibilità, per poter gestire richieste eterogenee e dinamiche, quali la disponibilità di applicazioni ottimizzate per sistemi operativi diversi e la possibilità di rendere disponibile una parte delle risorse in un ambiente Grid. Veniva altresì richiesto che il sistema garantisse un'adeguata espandibilità, sia come numero di nodi sia come possibilità di aggiungere schede ai nodi installati.

La scelta è caduta su un sistema di nodi biprocessori con tecnologia blade, sviluppati dalla medesima Exadron, con processori AMD Opteron 2,2 GHz *dual-core* [3] con 2 GB di RAM per *core*. L'utilizzo di server blade unito al processore *dual-core*, consente un'elevata compattezza, congiuntamente a un più ridotto consumo energetico. Il sistema si compone al momento attuale (Fig. 2) di 10 cestelli ripartiti su due armadi, mentre un terzo armadio contiene le apparecchiature di *storage* e di interconnessione (Fig. 3). Ogni cestello è composto di sette nodi, sei di dimensioni standard e uno più grande, tale da consentire l'installazione di schede aggiuntive (Fig. 4). In totale si compone quindi di 280 *core*, per una potenza di picco di 1,23 TFlop/s.

Per quanto riguarda l'interconnessione, i nodi sono dotati di interfaccia gigabit ethernet per la gestione e la connessione con lo *storage*, e di scheda Infiniband 4X [4] per il *message passing*. Infiniband è una delle tecnologie ormai più diffuse per le interconnessioni ad alte prestazioni, capace di 10 Gbit/s di banda passante con latenze di circa 2 μs.

Fig. 2 – Visione di insieme del sistema Michelangelo

Per quanto riguarda lo spazio disco, i cestelli sono connessi con una interconnessione Fibre Channel a 2 Gbit/s a due unità *storage*, una Nexsan Ataboy 2 [5] dotata di 14 dischi da 400 GB ciascuno, e una Nexsan Atabeast [5] dotata di 42 dischi da 500 GB ciascuno, entrambi in configurazione RAID per garantire la protezione dei dati. Lo spazio disco totale ammonta quindi a 26 TB.

Fig. 3 – Dettaglio di uno degli armadi di calcolo e dell'armadio di servizio di michelangelo.

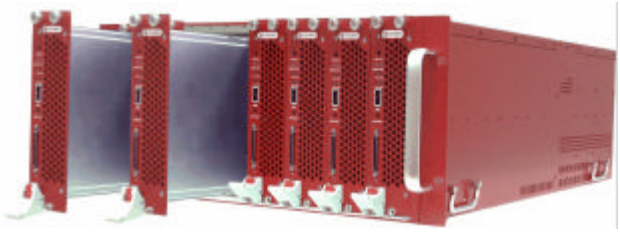

Fig. 4 – Dettaglio di uno dei cestelli blade.

L'infrastruttura è tale da garantire un raddoppio del numero dei nodi del sistema senza ricorrere a costose sostituzioni di componenti installati, e i nodi a doppia larghezza potranno ospitare una scheda programmabile di tipo Field Programmable Gate Array (FPGA), in corso di sviluppo presso uno dei partner del progetto, capace di produrre eccellenti prestazioni su specifici algoritmi.

Inoltre, la facilità di gestione sistemistica del sistema è garantita dall'utilizzo di un sistema KVMoIP (Keyboard Video Mouse over Internet Protocol), che consente un singolo e agevole punto di accesso alla console di tutti i nodi anche via web.

La configurazione software

Per quanto riguarda la configurazione software, si è deciso di replicare la stessa struttura già utilizzata con successo sul cluster *golgi.cilea.it* [6], ove i nodi di calcolo hanno una configurazione *diskless* basata su UnionFS [7], che garantisce al contempo la robustezza e la flessibilità del sistema. Un tale sistema infatti elimina i tempi di non disponibilità dei nodi dovuti a problemi relativi al disco o al bus SCSI, una frazione che l'esperienza maturata sul cluster *avogadro.cilea.it* [8] ha dimostrato essere elevata. Questa soluzione garantisce la possibilità di agevoli variazioni nella configurazione dei nodi.

Così, come sul sistema *golgi,* la maggioranza dei nodi installa un sistema operativo Linux CentOS [9]. Fa eccezione un numero limitato e variabile di nodi dotati di SUN Solaris 10 [10], per permettere l'utilizzo di particolari applicazioni, o Scientific Linux [11], per soddisfare i requisiti del *middleware* Grid del Progetto EGEE (Enabling Grids for E-sciencE) [12]. Altri sistemi operativi, quali Microsoft Windows, potranno essere installati su richiesta.

Per un'efficace gestione degli *storage* si è scelto GFS (Global File System) [13] come *cluster file system*. GFS permette a un insieme di computer di utilizzare simultaneamente un'area disco: legge e scrive sul *device* come se fosse un *filesystem* locale, ma è dotato di un sistema di *lock* per permettere ai vari client di coordinare le operazioni di I/O in maniera tale da mantenerne la consistenza. I *filesystem* definiti sugli *storage* sono connessi mediante GFS ai nodi dotati di scheda FibreChannel, e da questi condivisi agli altri nodi del proprio cestello via NFS (Network File System).

La configurazione del kernel Linux per i nodi di calcolo si è rivelata piuttosto complessa, perché le varie componenti del sistema avevano richieste tra loro incompatibili, come schematizzato in Fig.5. Per esempio, i *driver* per la connessione Infiniband richiedevano un *kernel* di versione inferiore alla 2.6.11, quando la gestione dei processori AMD *dual-core* era possibile solo a partire dalla versione 2.6.12. È stato quindi necessario procedere alla costruzione di un *kernel* ad hoc, partendo dalla versione 2.6.16.16 vanilla e andando a modificare i vari moduli necessari affinché potessero lavorare con il medesimo.

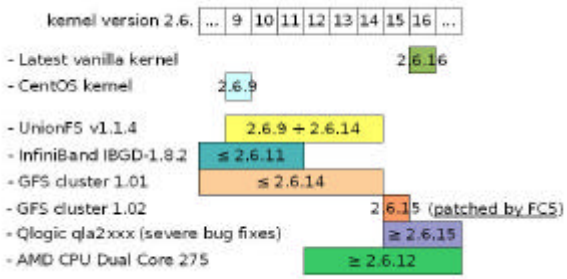

Fig. 5 – Tabella di compatibilità dei componenti software rispetto al kernel Linux.

Configurazione di un sistema diskless

La struttura di un cluster *diskless* è piuttosto complessa e si fonda su diverse componenti. Le Figg. 6 e 7 schematizzano le varie fasi del *boot* di un nodo *diskless*.

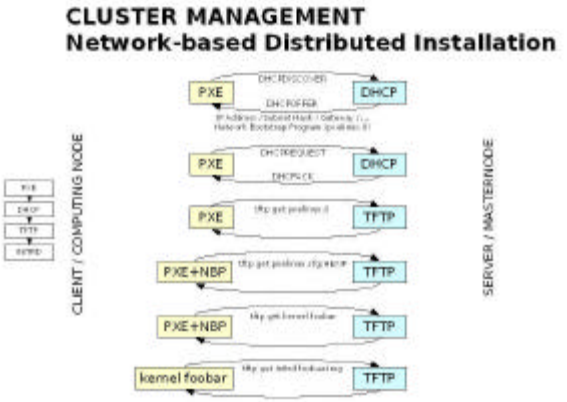

Fig. 6 – Schema che illustra il caricamento del kernel mediante PXE su un nodo diskless.

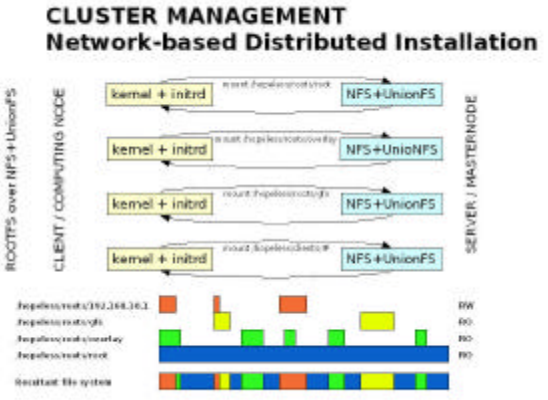

Fig. 7 – Schema che illustra il caricamento delle cartelle di sistema mediante UnionFS su un nodo diskless.

Il nodo esegue il *boot* via rete mediante il sistema PXE (Preboot eXecution Environment) [14], che innanzitutto richiede al DHCP server un opportuno indirizzo IP, che usa per richiedere il trasferimento via tftp del *kernel* e dei file che permettono di mettere in piedi l'ambiente operativo. Il *filesystem* di *root* (rootfs) viene montato via NFS e la struttura delle cartelle di sistema viene costruita sovrapponendo diversi livelli globali o specifici per il nodo mediante il sistema UnionFS. Rispetto a una configurazione in cui sul nodo master abbiamo singole immagini per ciascun nodo, questo sistema garantisce la coerenza del sistema e la facilità di mantenimento delle componenti perché basta effettuare una singola modifica su una cartella contenuta sul master e questa sarà immediatamente disponibile a tutti i nodi di calcolo.

Il controllo sistemistico del sistema è poi affidato a un insieme di strumenti opensource, come Ganglia [15], o realizzati in proprio.

La descrizione dell'ambiente di lavoro per gli utenti e delle applicazioni installate verrà affidata a un futuro articolo [16].

L'ambiente Grid

Gli scopi del Progetto LITBIO prevedono che il sistema possa, almeno in parte, essere disponibile in un ambiente di griglia computazionale, per essere concretamente parte di un network europeo della ricerca bioinformatica. La struttura flessibile del sistema Michelangelo ne permette l'inserimento con facilità.

In particolare, il cluster è stato configurato per integrarsi con un Grid basato su *middleware* LCG2/gLite [17], quale quello definito dal progetto europeo EGEE [12], con il seguente meccanismo:

- Uno dei nodi di calcolo viene configurato come Computing Element (CE), opportunamente dotato di un sistema operativo (Scientific Linux 3.0.6) che supporti l'installazione del middleware.
- Il Computing Element si interfaccia al sistema di code del nodo master.
- Su ogni nodo del *cluster* è installato il Worker Node (WN) *middleware* necessario per gestire *job* provenienti dal Computing Element.

Conclusioni

Il calcolatore appena installato si pone come una delle più rilevanti realtà europee dedicate esclusivamente alla bioinformatica, con notevoli punti di forza derivanti dalla sua scalabilità e flessibilità nella configurazione.

Inoltre, il sistema nasce dalla lunga esperienza CILEA nel campo dei cluster e dei

calcolatori ad alte prestazioni, e può essere integrato con il resto del nostro parco macchine. Allo scopo di testare la funzionalità della macchina e di saggiarne le potenzialità si è provveduto a un test LINPACK, combinandolo con i cluster golgi e avogadro. Il test pre-supponeva delle sfide tecnologiche di rilievo, vista l'eterogeneità sia dal punto di vista dell'architettura dei processori, sia da quella delle tecnologie di interconnessione. Il sistema combinato, che ha una potenza di picco di 3770 GFlop/s, ha ottenuto una potenza sostenuta di 2130 GFlop/s, che ha permesso il ritorno del CILEA nella classifica TOP500 [18] di giugno 2006 alla posizione 464. La scommessa può dirsi vinta.

Bibliografia

[1] LITBIO
URL: http://www.litbio.eu

[2] EXADRON
URL: http://www.exadron.com/

[3] AMD
URL: http://www.amd.com

[4] Infiniband Trade Association
URL: http://www.infinibandta.org/

[5] NEXSAN
URL: http://www.nexsan.com/

[6] C. Arlandini, "GOLGI: un cluster Opteron per il CILEA", Bollettino del CILEA n. 101 Aprile 2006

[7] UnionFS
URL: http://www.am-utils.org/project-unionfs.html

[8] C. Arlandini, "AVOGADRO: il CILEA oltre il muro del TeraFlops", Bollettino CILEA, 91 Febbraio 2004

[9] CentOS
URL: http://www.centos.org/

[10] SUN Solaris
URL: http://www.sun.com/software/solaris/

[11] Scientific Linux
URL: http://www.scientificlinux.org/

[12] EGEE Project
URL: http://www.eu-egee.org/

[13] GFS
URL: http://sources.redhat.com/cluster/gfs/

[14] PXE
URL: http://www.pxe.ca/

[15] Ganglia Monitoring System
URL: http://ganglia.sourceforge.net/

[16] C. Arlandini, M. Marchisio, P. Ramieri, Bollettino CILEA, in prep.

[17] gLite middleware,
URL: http://glite.web.cern.ch/

[18] TOP500
URL: http://www.top500.org