AVOGADRO: il CILEA oltre il muro del TeraFlops

Claudio Arlandini
CILEA, Segrate

Abstract
Il CILEA, in collaborazione con le Università di Milano, Milano Bicocca, e Politecnico di Milano, potenzia il suo parco macchine per il calcolo parallelo ad alte prestazioni con un cluster di 128 nodi biprocessori Intel Xeon 3.06GHz, fornito dalla Exadron, Divisione High Performance Computing di Eurotech Spa. Con questa acquisizione il CILEA diversifica la sua offerta e si conferma tra le più importanti realtà del supercalcolo europeo.

Keywords: Hardware, Supercalcolo, Calcolo Parallelo.

Come annunciato sul bollettino precedente [1] nei primi giorni di gennaio 2004 è stato installato al CILEA il nuovo supercalcolatore che andrà a potenziare l'offerta a disposizione degli utenti nel settore del calcolo ad alte prestazioni. Questo articolo vuole essere una scheda tecnica più dettagliata del nuovo server.
Il Consorzio ha deciso di diversificare maggiormente la propria offerta, che ricordiamo comprendere un cluster di macchine HP con processori a 64 bit [2] ed una macchina vettoriale NEC [3], riconoscendo le ottime prestazioni raggiunte dai più recenti processori 32bit di Intel e AMD, la stabilità raggiunta dal sistema operativo Linux e delle infrastrutture software di clustering, e della disponibilità crescente di software tecnico-scientifico per questa piattaforma. L'interesse del mondo del calcolo ad alte prestazioni per i cluster di nodi con processori a 32 bit è infatti testimoniato dalla crescita esponenziale di macchine di questo tipo che si riscontra nella classifica TOP500 [4] negli ultimi due anni, tanto da superare ormai il 40% del totale. Sebbene per molte applicazioni calcolatori di tipo SMP o vettoriali rimangano insostituibili, non fosse altro che per la capacità di gestire efficacemente quantità di memoria molto superiori, è pur vero che il rapporto prezzo/prestazioni di un cluster di componenti COTS (Commodities Off-the-Shelf) è molto invitante.

Il CILEA è però andato oltre questo paradigma e offre ai suoi clienti quanto di meglio questa tecnologia può offire, basandosi su processori Intel Xeon dell'ultima generazione, e di una rete di interconnessione ad alta velocità e bassa latenza Myrinet 2000. I benchmark Linpack [5] effettutati finora hanno toccato prestazioni di 1,084 TeraFlops, che collocheranno questa macchina nella parte alta della prossima classifica TOP500.
Il server è stato dedicato ad un esimio scienziato italiano, nella tradizione inaugurata con il precedente server SuperDome. L'onore è toccato questa volta ad Amedeo Avogadro (1776-1858), scienziato di origine torinese (fig.1), autore di scoperte fondamentali nel campo della chimica e noto soprattutto per la legge e il numero che portano il suo nome. Scopo di questa dedica è non solo un piccolo omaggio a questo grande pensatore, ma anche un augurio che il nuovo server CILEA sia strumento importante per nuove scoperte scientifiche ed innovazioni tecnologiche.

Figura 1 - Ritratto di Lorenzo Romano Amedeo Carlo Avogadro, conte di Quaregna e Cerreto.

Scheda Tecnica

Il Comitato Tecnico del CILEA ha scelto come fornitore del sistema, dopo approfondita selezione, la Exadron, Divisione High Performance Computing di Eurotech SpA, con sede ad Amaro (UD) [6].

Il cluster si compone di 128 nodi biprocessori, server SuperMicro con processori Intel Xeon 3.06GHz su motherboard "Super X5DPA-8GG". Tra le caratteristiche tecniche del processore ricordiamo 512 KB di cache, system bus a 533 MHz, e tecnologia Hyper-Threading. Questa tecnologia fornisce un parallelismo a livello di thread (TLP) su ogni processore, permettendo quindi un maggior utilizzo delle risorse. E' quindi una forma di multi-threading simultaneo dove threads multiple di un'applicazione software possono essere eseguite contemporaneamente su un processore. Questo è ottenuto duplicando l'architettura del processore, pur mantenendo un solo insieme di risorse di esecuzione. Il risultato è quindi quello di ottenere un generale incremento di prestazioni per applicazioni multitasking. La scelta tra attivazione o meno di Hyper-Threading a livello di singolo nodo è facilmente modificabile, consentendo quindi di lavorare nella condizione ottimale per l'applicazione in esecuzione.

Ogni nodo è dotato di disco interno SCSI da 36 GB a 10000 rpm. Il nodo ha 4 GB di memoria RAM ed ha due interfacce di rete: una Gigabit Ethernet per l'amministrazione e l'accesso allo storage condiviso, ed una Myrinet 2000 per il message passing.

Il sistema è contenuto in quattro armadi 42U. Per ciascuno di questi è stato predisposto un sistema di storage collegato via SCSI con uno dei nodi, capace di 1.2 TB. In totale quindi gli utenti potranno disporre di ben 6.4 TB di spazio disco.

Le unità di storage sono Nexsan ATABoy, contenenti dieci dischi IDE da 120 GB. Il sistema ha a bordo una cache da 128 MB. I dischi sono stati organizzati in una struttura RAID 5, per garantire la conservazione dei dati anche in caso di guasto ad uno dei dischi. Su queste unità di storage sono state costruite le aree degli utenti e lo scratch condiviso. Attualmente la condivisione di queste aree a tutti i nodi avviene tramite NFS, ma sono in corso valutazioni di infrastrutture più performanti quali il file system Lustre [7], o il Parallel Virtual File system (PVFS) [8].

Il sistema operativo adottato è Linux Red-Hat 9, con kernel 2.4.23. Sono in corso test per verificare la stabilità dei kernel di generazione 2.6.x per il futuro upgrade. Questa generazione di kernel consente tra l'altro una gestione migliore dell'hyperthreading. L'accordo tra CILEA ed Exadron prevede infatti non solo la fornitura del sistema ma un'effettiva partnership per un costante upgrade hardware e software, allo scopo di ottenere un continuo incremento delle prestazioni.

Il cluster è amministrato e gestito tramite *OSCAR* (Open Source Cluster Application Resources) [9], un pacchetto open-source consistente in un insieme altamente integrato di strumenti per il cluster computing, quali esecuzione contemporanea di comandi su più nodi, spostamento di files, gestione degli utenti, monitoraggio, e così via. Il sistema di gestione delle code è *Torque* [10], un sistema basato su *OpenPBS* 2.3.12 [11] con innovazioni rilevanti per quanto riguarda la scalabilità, l'usabilità, e la tolleranza ai guasti. A questo è stato associato *Maui* [12], uno scheduler avanzato che permette una più efficace politica di gestione delle risorse e "fairshare" tra gli utenti.

E' comunque in corso di valutazione il passaggio a LSF, che permetterebbe un accesso integrato anche alle altre risorse di carico già presenti.

Ognuno dei processori è capace di 6.12 GFlops di picco. Abbiamo misurato un valore di 1084 GFlops di potenza sostenuta (efficienza 70%), il che significa che con l'inserimento della nuova macchina la potenza di calcolo disponibile al CILEA è quintuplicata. I primi giorni del mese di gennaio sono stati dedicati allo studio e alla ottimizzazione delle prestazioni, mentre

l'apertura a tutti gli utenti è avvenuta il 19 gennaio 2004.
Nella seguente tabella riportiamo le caratteristiche salienti del calcolatore.

Cluster Exadron	
CPUs (nodi)	256 (128)
Processore	Intel Xeon 3.06 GHz
RAM per nodo	4 GB
Spazio disco	6.4 TB
OS	Linux Red-Hat 9
Kernel	2.4.23
Interfaccie di rete	2 (Myrinet2000, gigabit ethernet)

Myrinet2000

La rete di interconnessione ad alte prestazioni destinata al message passing per applicazioni MPI multiprocessore è Myrinet 2000, uno dei sistemi più noti ed utilizzati per cluster di questo tipo, prodotta dalla Myricom [13].
Questo tipo di interconnessione si caratterizza per l'elevato throughput, 2Gbit/s full duplex, unito ad una bassa latenza, tipicamente 7 μs per messaggi piccoli. I nodi sono collegati tramite un singolo switch a 128 porte (fig.2), che consente la costruzione di un'architettura di rete tipo "fat-tree", ottimizzando la comunicazione tra due nodi qualsiasi del cluster.

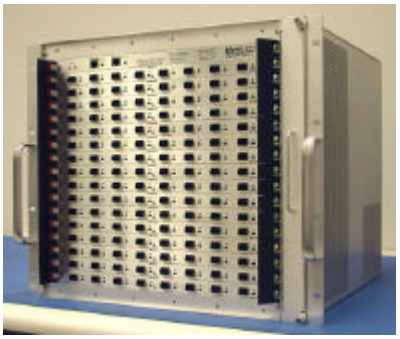

Figura 2 - Switch Myrinet a 128 porte in fibra.

Myricom fornisce anche librerie ad alte prestazioni destinate ad essere integrate in MPICH [14], la più nota delle implementazioni open-source delle librerie MPI, il cui sviluppo è curato dal prestigioso Argonne National Laboratory, che permettono una comunicazione tra nodi senza passare attraverso lo stack TCP-IP, con conseguente diminuzione della latenza nelle comunicazioni e conseguente aumento delle prestazioni [15].

Modalità di utilizzo

La configurazione attuale prevede che gli utenti accedano direttamente al cluster collegandosi al nodo di front-end, di indirizzo avogadro.cilea.it, in modalità SSH v.2. L'utilizzo degli altri nodi è consentito unicamente attraverso la modalità batch, usando il sistema di gestione di code *Torque*. In particolare è definita una coda *default* il cui compito è smistare i job a seconda delle risorse richieste in quattro code, aventi priorità inversamente proporzionali alla quantità di risorse richieste: *small, medium, long e verylong*.
Per i principali pacchetti applicativi sono stati preparati opportuni script ottimizzati per il corretto lancio in coda.
L'esecuzione di un job può avvenire sia a linea di comando, con *qsub*, sia con interfaccia grafica *xpbs*, che consente anche in maniera intuitiva il controllo e l'interruzione dei lavori. Seguirà un articolo dettagliato sull'utilizzo del sistema da parte degli utenti.
Per ora rimandiamo gli interessati all'assistenza dell'autore Dott. Claudio Arlandini (arlandini@cilea.it) oppure del Dott. Maurizio Cremonesi (cremonesi@cilea.it).
Attualmente sono disponibili i seguenti pacchetti:

Ambienti di sviluppo:
(referente Dott. Maurizio Cremonesi)

Package		Release
GCC		3.2.2-5
Intel Fortran Compiler		8.0
Intel C++ Compiler		8.0
Intel VTune		2.0
Intel MKL		7.0
MPICH		1.2.5
MPICH-GM		1.2.5.11
LAM-MPI		7.0

Analisi strutturale:
(referente Dott. Maurizio Cremonesi)

Package		Release
ABAQUS		6.4
ANSYS		7.0
LS-DYNA		960
MSC suite	NASTRAN	2001
	PATRAN	2004
RADIOSS		

Termo-fluidodinamica:
(referente Dott. Paolo Ramieri)

Package		Release
FLUENT suite	FLUENT	6.1.22
	FIDAP	8.7.2
	Gambit	2.1
	Tgrid	3.5
STARCD suite	Starcd/hpc	3.150a
COMET		

Elettromagnetismo:
(referente Dott. Maurizio Cremonesi)

Package		Release
ANSYS		7.0

Chimica:
(referente Dott. Maurizio Cremonesi)

Package		Release
GROMACS		3.2
TURBOMOLE		5.6
CPMD		3.7
GAMESS		December 12, 2003

Per avere ulteriori informazioni sugli applicativi disponibili rivogersi ai referenti:
Dott. Maurizio Cremonesi (cremonesi@cilea.it)
Dott. Paolo Ramieri (ramieri@cilea.it)

Bibliografia

[1] A. Cantore, "Un nuovo supercalcolatore al CILEA: un cluster da oltre un TeraFlops", Bollettino del CILEA, 90, Dicembre 2003
[2] C. Arlandini, "GALILEO: un nuovo server HP SuperDome per il calcolo parallelo al CILEA", Bollettino del CILEA, 81, febbraio 2002
[3] F. Bonini, A. Mattasoglio, "Nuovo calcolatore vettoriale del CILEA: Nec SX-5/4s", Bollettino del CILEA, 76, Febbraio 2001
[4] TOP500, URL: http://www.top500.org
[5] LINPACK Benchmark, URL: http://www.netlib.org/linpack/
[6] EXADRON, URL: http://www.exadron.com
[7] LUSTRE filesystem, URL: http://ww.lustre.org
[8] PVFS filesytem, URL: http://www.parl.clemson.edu/pvfs/
[9] OSCAR, URL: http://oscar.sourceforge.net/
[10] Torque Resource Manager, URL: http://www.supercluster.org/projects/torque/
[11] OpenPBS, URL: http://www.openpbs.org/
[12] Maui scheduler, URL: http://www.supercluster.org/maui/
[13] Myricom, URL: http://www.myri.com
[14] MPICH, URL: http://www-unix.mcs.anl.gov/mpi/mpich/
[15] Prestazioni di Myrinet 2000, URL: http://www.myri.com/myrinet/performance/index.html