Centro Regionale di Competenza
Analisi e Monitoraggio del Rischio Ambientale

Istituto Nazionale per la Fisica della Materia – SGD Campania

CluES

Cluster for Environmental Simulations

Manuale tecnico a cura di:

Lucilla de Arcangelis
Antonio De Candia
Francesco Maria Taurino

Centro Regionale di Competenza
Analisi e Monitoraggio del Rischio Ambientale

Istituto Nazionale per la Fisica della Materia – SGD Campania

CluES

Cluster for Environmental Simulations

Manuale tecnico a cura di:

Lucilla de Arcangelis
Antonio De Candia
Francesco Maria Taurino

Centro Regionale di Competenza
Analisi e Monitoraggio del Rischio Ambientale
Polo delle Scienze e delle Tecnologie
Dipartimento di Scienze Fisiche
C/o Facoltà di Ingegneria - Via Nuova Agnano, 11 - III Piano
80125 - Napoli - Italy
www.amra.unina.it
ambiente@na.infn.it
Telefono +39 081 76-85125/124/115
Fax. +39 081 76-85144

Autori
Lucilla de Arcangelis, Antonio De Candia, Francesco Maria Taurino
Istituto Nazionale per la Fisica della Materia – SGD Campania

Coordinamento editoriale
doppiavoce
www.doppiavoce.it

Indice

Introduzione

Negli ultimi venti anni si è progressivamente sviluppato l'interesse del mondo scientifico per l'utilizzo di strumenti di calcolo numerico sia dal punto di vista dell'hardware, con la messa a punto di calcolatori sempre più potenti, sia di quello del software, con lo sviluppo di algoritmi complessi. La ragione di tale crescente successo è il grande impatto che il metodo di studio numerico ha avuto sia nell'ambito della ricerca fondamentale sia in quello della ricerca applicata. Metodi di calcolo numerico messi a punto in ambiti di ricerca di base si sono, infatti, rivelati di grandi potenzialità se applicati allo studio di problemi rilevanti nell'ambito dell'analisi del rischio ambientale, quali risposta meccanica dei materiali, dinamica dei terremoti, frane in mezzi granulari o formazione di inquinanti gassosi e particellari in reattori di combustione.
Lo sviluppo delle attività di calcolo in ambito scientifico è stato senza dubbio sostenuto dal perfezionamento di macchine di calcolo sempre più competitive. Negli ultimi dieci anni la metodologia degli operatori del campo è evoluta dal calcolo vettoriale su computer main-frame di grande potenza, verso il calcolo parallelo su cluster di processori. Questa tecnologia, oramai molto diffusa, permette di realizzare con costi moderati uno strumento di calcolo efficiente e avanzato.
D'altra parte, lo sviluppo estremamente veloce delle tecnologie informatiche fa sì che ogni apparecchiatura di calcolo, per quanto competitiva alla nascita, diventi nel giro di pochi anni superata, rendendo quindi praticamente impossibile l'acquisizione di un supporto per il calcolo avanzato che sia efficiente sui tempi lunghi.
Per fare fronte a questa esigenza, l'orientamento degli operatori del settore in anni recenti si è rivolta verso la tecnologia GRID. Quest'ultima organizza in un'unica rete più macchine di calcolo distribuite sul territorio, permettendo l'accesso e suddividendo l'onere del calcolo su ciascuna di esse come macchina dell'utente. Questa tecnologia consente quindi al singolo operatore di usufruire di un gran numero di risorse di calcolo, scelte in base alla disponibilità del momento, senza essere penalizzato

da problemi hardware o necessità di upgrade della macchina locale.
L'attenzione per la tecnologia GRID è in fase di rapida diffusione in Europa e negli Stati Uniti. Negli ultimi anni l'INFN e il CNR hanno ricevuto notevoli finanziamenti nazionali ed europei per sviluppare in Italia il progetto GRID. Tali attività, basandosi sulla rete GARR, sono principalmente centrate sullo sviluppo del middleware, cioè del software di gestione del GRID. Questi strumenti software devono poter gestire l'accesso, la sicurezza, l'utilizzo delle varie risorse e il monitoraggio della rete GRID. Al tempo stesso l'INFN ha in atto progetti che si occupano di integrare il middleware europeo con quello degli Stati Uniti, rendendo compatibili le due reti con l'integrazione dei protocolli e creando la prima infrastruttura GRID intercontinentale. Tali attività sono sviluppate in sinergia con i maggiori centri di ricerca europei nel campo della fisica delle alte energie e coordinate dal CERN. L'INFM partecipa con questi Enti al progetto Campus-Grid presso l'Università di Napoli Federico II con il cluster CluES del Centro Regionale di Competenza AMRA.
La realizzazione di tale progetto può avere un grande impatto sullo sviluppo delle attività di ricerca. I diversi metodi numerici utilizzati si gioveranno in maniera differente di una struttura GRID. Metodi Monte Carlo, ad esempio, hanno una semplice distribuzione del calcolo parallelo simulando diverse configurazioni di disordine su processori diversi. In tal caso la potenza del singolo processore determina l'efficienza del calcolo. Il metodo agli elementi finiti, d'altra parte, suddivide l'intero sistema su più processori e quindi richiede l'utilizzo di un gran numero di nodi di calcolo. Questi due esempi estremi mostrano come per realizzare calcolo avanzato la soluzione GRID mette a disposizione soluzioni flessibili alle richieste del singolo utente.
Su tale supporto di calcolo sarà quindi possibile organizzare un'offerta di modellistica numerica diversificata. Come esempio concreto, le attività già avviate presso l'Unità di Napoli dell'Istituto Nazionale per la Fisica della Materia (INFM) hanno come obiettivo la messa a punto di ampia gamma di modelli numerici per la simulazione di problemi nel campo dei materiali o sistemi disordinati, di forte interesse industriale ma anche di interesse per il rischio ambientale: frattura di materiali eterogenei, automi cellulari applicati alla propagazione della lava

durante le eruzioni, proprietà statistiche dei cataloghi sismici, dinamica delle frane su pendii e così via. Tali strumenti sono altresì applicabili allo studio di un'ampia gamma di problemi fisici rilevanti. Con tali attività AMRA si offre quindi come possibile partner a Industrie o Enti che intendono presentare progetti e richiedono un supporto di modellistica numerica.

Descrizione tecnica

CluES è un cluster di tipo Beowulf per calcolo ad alte prestazioni, che utilizza computer indipendenti combinati in un unico sistema facendo uso di appositi software e apparati di rete. Il cluster è costituito da 99 macchine biprocessore, da una rete privata standard a 100/1000 mb/s, da una rete ad alta velocità e bassa latenza, da un sistema di storage NAS e da diversi apparati per il controllo remoto (Figura 1).
I nodi del supercomputer sono dei server in formato rack da una unità (1 U) prodotti dalla società americana APPRO [1], disposti fisicamente in 4 rack da 19 pollici per 42U.
Ogni nodo è così configurato:

– scheda madre Tyan Thunder S2469, con chipset AMD-760 MPX;
– 2 processori AMD Athlon [2] MP 2800+, funzionanti a 2133 Mhz, con 512 KB di cache di secondo livello;
– 4 GB di ram in formato ECC registered;
– un disco fisso Ultra ATA da 80 GB;
– due schede di rete, a 100 e 1000 mb/s.

Il cluster può contare quindi su 198 processori e 396 GB di memoria. Un singolo server può erogare una potenza di calcolo di circa 900 SpecINT [3] e 750 SpecFP.
I 99 nodi sono collegati fra loro attraverso 3 switch Allied Telesyn [4] AT8748XL, con 48 porte fast ethernet e 2 porte di uplink in gigabit ethernet. Questo sistema di connessione viene utilizzato principalmente per l'installazione e la manutenzione del cluster, e come rete di comunicazione di backup per lo scambio dei dati dei programmi nel caso di malfunzionamenti della rete a bassa latenza.

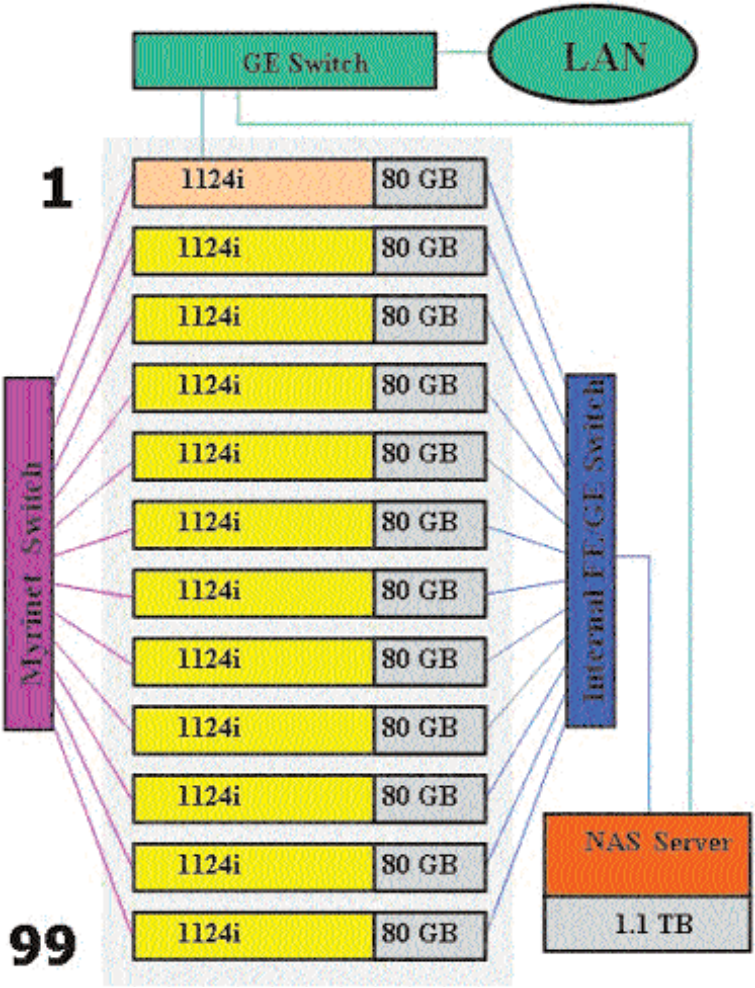

Fig. 1. Schema del CluES.

Per la rete a bassa latenza e alta velocità è stato scelto il sistema Myrinet, della società Myricom [5] (Figura 2). In ogni nodo è installata una scheda in fibra ottica modello 2000 PCI64B, collegata a un unico switch, l'M3-E128, con 104 porte da 2,5 gigabit al secondo e latenza media di circa 6/7 microsecondi.

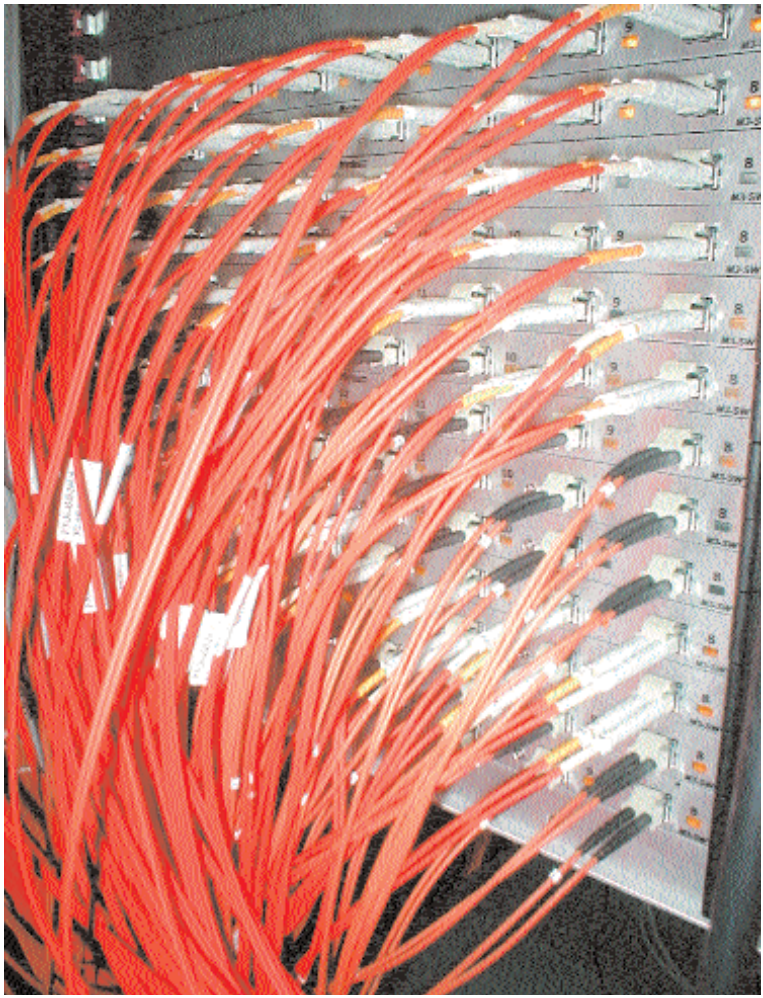

Fig. 2. Il sistema Myrinet.

Questa rete viene utilizzata per il message passing dei programmi di calcolo e simulazione.
Il sottosistema di gestione prevede l'utilizzo dei terminal server Cyclades [6] AlterPath ACS, che danno la possibilità di un accesso in console su ogni nodo attraverso le porte seriali, e

degli switch KVM (Keyboard Video Mouse) KeeMux della NTI [7], per i casi di emergenza (Figura 3).
Per lo storage, in questa prima fase, è stato installato un NAS (Network Attached Storage) della Quantum: lo SnapServer [8] 14000 (Figura 3). Questo sistema è un appliance capace di esportare verso la rete uno spazio disco di circa 1,1 TB, attraverso due schede gigabit dedicate. Il volume è formato da 12 dischi da 120 GB configurati in raid 5, uno dei quali utilizzato come hot spare in caso di guasti.
Il backup è affidato alla libreria IBM TotalStorage [9] Ultrium 3582, che utilizza tape in formato LTO 2 (Linear Tape Open) da 200 GB in formato non compresso. Questo modello è dotato di un drive e di 24 slot per cassette, con una capacità totale di 4,8 GB non compressi/9,6 GB compressi.
Il software di base installato sul cluster è Scientific Linux [10] 3, una distribuzione di linux basata sugli RPM free della RedHat [11] Enterprise Linux 3, sviluppata da grandi laboratori di ricerca come il FermiLab e il Cern.
Questa è ora la distribuzione di riferimento per i grossi progetti di ricerca scientifica. Semplifica l'installazione, la gestione e il monitoring del cluster e comprende una serie di software di base, come la distribuzione linux stessa e il sistema di replica delle installazioni basato sul kickstart della RedHat, e alcuni software opzionali e configurazioni speciali, come openafs [12] e il redirect della console su porta seriale.
L'installazione comprende inoltre i tool di sviluppo e compilatori standard, come il gcc e g77, la libreria MPICH [13] con supporto per la rete ethernet e Myrinet, il sistema di monitoring Ganglia [14].
È stato installato anche il sistema di code Torque [15], derivato da OpenPBS. In futuro verrà inoltre installato il middleware Grid di LCG [16].
Lo schema concettuale del cluster comprende un nodo master, chiamato “frontend”, e diversi nodi di calcolo (compute node). Gli utenti del sistema possono collegarsi solo e unicamente al frontend, su cui effettuano operazioni di editing e compilazione, e da cui poi sottomettono i job di calcolo. Questi ultimi vengono collocati nell’apposita coda di esecuzione, con priorità e caratteristiche diverse in base alle esigenze e al gruppo di appartenenza dell’utente.

Fig. 3. Alter Path ACS, Switch KVM, unità di backup IBM e Snap Server.

Possibili applicazioni

Il cluster CluES può essere usato per eseguire calcolo scientifico di diverso tipo:
– simulazioni Monte Carlo;
– simulazioni di dinamica molecolare;
– risoluzione di equazioni differenziali a derivate parziali
– studio della turbolenza nella dinamica dei fluidi
– determinazione della struttura elettronica di molecole
– simulazione dello scorrimento del traffico di veicoli.

La gran parte di queste applicazioni può essere rilevante nello studio dei problemi di rischio ambientale. Ad esempio, la Figura 4 mostra la distribuzione di sforzi locali in un mezzo eterogeneo fratturato, ottenuta tramite simulazioni Monte Carlo su reticolo. Una tensione esterna è applicata in alto e in basso ai bordi del sistema. I tratti bianchi spessi individuano gli elementi mesoscopici fratturati mentre il colore individua l'intensità dello sforzo.

Un cluster di computer interconnessi permette di eseguire in tempo ragionevole calcoli che richiederebbero tempi proibitivi su un normale computer seriale, mediante la parallelizzazione del programma. Questo vuol dire che un certo numero di processori lavora in parallelo, e ciascuno di essi esegue una parte dei calcoli che devono essere svolti.
La parallelizzazione può avvenire fondamentalmente in due diverse modalità.

1. La prima è quella che viene usata più spesso nel campo della fisica statistica, ad esempio nel calcolo delle proprietà termodinamiche di un fluido, o di un materiale granulare, in cui le quantità da misurare presentano delle fluttuazioni statistiche. In tal caso è necessario simulare un numero elevato di copie del sistema, soggette a rumore termico diverso, ed eseguire delle medie di insieme delle quantità di interesse. Su un computer parallelo si può allora simulare una copia del sistema su ciascun processore, con un fattore di "speed up" pari proprio al numero di processori utilizzati, che è anche

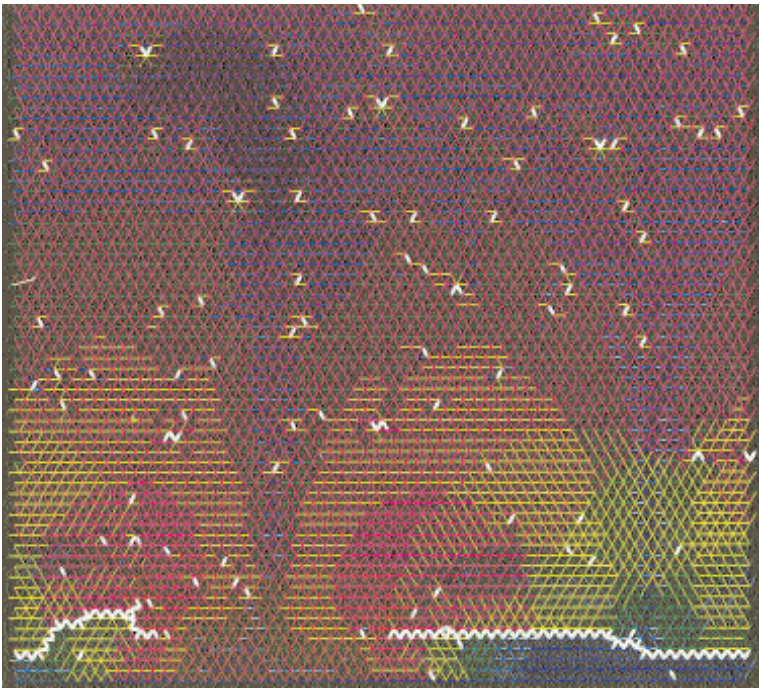

Fig. 4. Simulazioni Monte Carlo su reticolo: distribuzione di sforzi locali in un mezzo eterogeneo.

ovviamente il massimo guadagno ottenibile. Questo tipo di calcoli paralleli viene a volte chiamato “embarrassingly parallel computations”, appunto perché si ottiene il massimo guadagno con il minimo sforzo.

2. La seconda modalità di parallelizzazione viene usata quando il problema è caratterizzato da un numero di variabili estremamente elevato. Esempi includono la risoluzione di un’equazione differenziale alle differenze finite, oppure i calcoli di fluidodinamica, che richiedono la scomposizione dello spazio in una griglia abbastanza fitta, ai cui nodi sono definite delle variabili di cui si vuole trovare il valore, come ad esempio la velocità o la pressione del fluido. In tal caso è conveniente assegnare a ogni processore una regione diversa dello spazio, in modo da ridurre il tempo complessivo di

calcolo. Diversamente dal caso precedente, in questo tipo di problemi diventa importante la comunicazione tra i diversi processori, per poter raccordare le soluzioni nelle diverse regioni. La necessità di scambiare informazioni tra i diversi processori può rendere il fattore di guadagno temporale sensibilmente minore del numero di processori impiegato.

Per ridurre questo problema, è importante avere una rete di connessione tra i diversi computer molto veloce. I parametri fondamentali che caratterizzano una rete di connessione sono il "data rate", ovvero il numero di bytes che si possono trasferire nell'unità di tempo, e la "latenza", ovvero il tempo minimo necessario per trasferire anche un solo byte di dati. Il cluster CluES è dotato di una rete di connessione tra i processori di tipo Myrinet, che è caratterizzata da un elevato "data rate" e insieme da una latenza molto bassa, per cui è particolarmente adatto per svolgere anche un tipo di calcolo parallelo che usa in maniera intensa lo scambio di messaggi tra i diversi processori.
Lo scambio di messaggi tra i processi può essere realizzato su CluES utilizzando la libreria MPI (Message Passing Interface), installata sul cluster sia nella versione per linguaggio C che in quella per Fortran. Il meccanismo fondamentale di comunicazione di MPI è la trasmissione di dati tra due processi, il "sender" che spedisce un certo pacchetto, e il "receiver" che lo riceve. Le funzioni di send e receive possono essere di due tipi:
1. blocking;
2. non-blocking.

Ad esempio, nel caso di un "blocking send", il processo che spedisce il pacchetto si blocca finché il receiver non ha ricevuto il messaggio. Poiché questo tempo di attesa potrebbe essere utilizzato per eseguire ulteriori calcoli, la libreria MPI rende disponibile una funzione "non-blocking" send. Quest'ultima si compone di due parti (analogamente per le operazioni di "receive"):
1. un'operazione di "posting" in cui si fa partire il pacchetto;
2. un'operazione di "test-for-completion" in cui si verifica se il pacchetto è stato effettivamente ricevuto.

A partire da queste operazioni basilari di comunicazione “point-to-point”, la libreria MPI implementa anche operazioni più complesse, in cui la comunicazione avviene tra un sottogruppo di processi di dimensione maggiore di due. Queste operazioni comprendono, ad esempio, la sincronizzazione di un gruppo di processi, il “broadcast” di un pacchetto di dati da un processo a tutti gli altri, oppure operazioni di somma, prodotto, minimo, massimo, ecc., di vettori elemento per elemento. Tali operazioni vengono effettuate mediante algoritmi che minimizzano il tempo di esecuzione [17].

Bibliografia

1. http://www.appro.com
2. http://www.amd.com
3. http://www.spec.org
4. http://www.alliedtelesyn.com
5. http://www.myri.com
6. http://www.cyclades.com
7. http://www.nti1.com
8. http://www.snapserver.com
9. http://www.storage.ibm.com
10. https://www.scientificlinux.org
11. http://www.redhat.com
12. http://www.openafs.org
13. http://www-unix.mcs.anl.gov/mpi/mpich/
14. http://ganglia.sf.net
15. http://www.supercluster.org
16. http://lcg.web.cern.ch/LCG/
17. http://www-unix.mcs.anl.gov/mpi/tutorial/index.html

Finito di stampare nel mese di novembre 2004
presso la LEGMA/Napoli