LA FORMULA 1 DELL'IT

Evoluzione, stato dell'arte e prospettive per l'HPC, il top del top nel calcolo computazionale

Feedback: redazione-cbr@edizionibig.it

di ***Giancarlo Magnaghi***

Il campo del supercalcolo, o High Performance Computing (HPC), rappresenta per l'IT quello che la Formula 1 rappresenta per le automobili o l'industria spaziale per l'aeronautica: una palestra in cui vengono affrontati i problemi di punta alle frontiere della tecnologia, che poi hanno ricadute anche sui mezzi utilizzati per le normali applicazioni commerciali e industriali.

Nella ricerca scientifica, la formulazione di nuove teorie presuppone la creazione di modelli matematici e la loro verifica basata su simulazioni numeriche. Le scienze computazionali utilizzano in modo sistematico metodi matematici numerici (supercalcolo) e strumenti informatici dotati di rilevante potenza di calcolo (supercomputer).

Un po' di storia

Nella seconda metà degli anni '70 fecero la loro comparsa i primi supercomputer vettoriali, che avevano una potenza di calcolo dell'ordine dei MFlops (1 MegaFlops = 10^6 Flops = 10^6operazioni in virgola mobile al secondo). Alla fine degli anni '90 si arrivò ai GFlops (1 GigaFlops = 10^9 Flops), poi ai TFlops (1 TeraFlops = 10^{12} Flops) e nel 2008 è stato raggiunto il traguardo del PFlops (1 PetaFlops = 10^{15} Flops).

Il primo supercomputer di rilevanza per la comunità scientifica fu, negli anni '70, il Cray-1, con una potenza di picco di 160 MFlops. In Italia, il primo supercomputer fu il Cray X-MP 12 installato nel 1984 al CINECA di Bologna.

I primi sistemi di supercalcolo degli anni '70 erano sistemi vettoriali mono-processore (come il Cray 1), poi nei primi anni '80 comparvero i sistemi vettoriali multi-processore a memoria condivisa SMP (SymmetricMulti Processor), come i sistemi Cray X-MP e Y-MP, a cui si aggiunsero nella seconda metà degli anni '80 i sistemi paralleli a memoria distribuita. A partire dai primi anni '90, si sono imposti i sistemi *Massively Parallel Processors* (MPP) - costituiti da cluster di nodi SMP basati su processori RISC o microprocessori con architettura X86 interconnessi da reti veloci come Infiniband - che oggi caratterizzano la maggior parte dei sistemi HPC sul mercato. La distinzione tra sistemi convenzionali e supercomputer diviene molto più sfumata. Infatti un sistema convenzionale opportunamente esteso può diventare un supercomputer, e la grande maggioranza dei supercomputer è costituita dai medesimi microprocessori commodity presenti nei server commerciali, nelle workstation e nelle console giochi.

La rapida affermazione dei cluster nel mondo HPC è dovuta proprio all'utilizzo di componenti commodity con prestazioni elevate e costi contenuti, che permettono di realizzare cluster che, a parità di potenza di calcolo, costano fino a 10 volte meno di un supercalcolatore tradizionale.

Il 98% dei 500 supecomputer più potenti del mondo sono già basati su cluster di processori multicore.

Architetture parallele e acceleratori

Le architetture parallele vengono classificate anche in base al loro modello di memoria: sistemi con *memoria con-*

divisa - SMP (Symmetric Multi Processors) e sistemi NUMA (Non-Uniform Memory Access) - e sistemi con *memoria distribuita*.

I sistemi con memoria condivisa sono generalmente dotati di memorie cache di terzo livello (L3) condivise tra i diversi processori in parallelo, oltre alle cache di secondo livello (L2) condivise a livello di microprocessore e quelle di primo livello (L1) a livello di singolo core.

Nei sistemi con memoria distribuita, i processori sono connessi tramite una rete di interconnessione, ogni processore può indirizzare direttamente solo la propria memoria locale e utilizza un protocollo a scambio di messaggi (*message passing*) per scambiare informazioni con gli altri nodi.

Nei sistemi HPC, si affiancano alle CPU vari tipi di *acceleratori hardware specializzati*, per aumentare le prestazioni, specialmente nelle elaborazioni sequenziali (*single thread*), che richiedono un'elevata "forza bruta" di elaborazione, e nei calcoli con un elevato livello di parallelismo. Il principio generale è quello di eseguire determinate operazioni in hardware/firmware anziché in software.

Esistono vari tipi di acceleratori.

Le **GPU** (Graphical Processing Unit) sono processori grafici di fascia alta che derivano dai modelli utilizzati nelle schede grafiche e nelle console giochi.

Gli **FPGA** (Field Programmable Gate Array), utilizzati da tempo nei componenti personalizzabili ASIC (Application-Specific Integrated Ccircuit), sono circuiti integrati general-purpose che possono essere riprogrammati anche dopo essere stati inseriti in un sistema. Assorbono poca potenza e utilizzano parallelismo e pipeline per fornire elevate prestazioni (fino a centinaia di GFlops).

Gli acceleratori ASIC SIMD (Single Instruction Multiple Data) sono costituti da centinaia di unità aritmetiche parallele in virgola mobile e raggiungono potenze di calcolo di un centinaio di GFlops.

Per la programmazione sono utilizzati tipicamente i linguaggi C, C++ e Fortran90, con librerie specifiche per la gestione del parallelismo, insieme a nuovi linguaggi paralleli object-oriented come Chapel, X10 e Fortress18.

Originariamente, i supercalcolatori erano dedicati alle applicazioni *number crunching* come modellazione chimica e fisica, prospezione geofisica per la ricerca degli idrocarburi, e applicazioni di punta della progettazione (fluidodinamica e calcolo strutturale). Il volume di dati prodotti a livello mondiale dal supercalcolo raddoppia ogni anno, poiché negli ultimi anni le simulazioni numeriche si sono estese ai campi più disparati, come simulazione biomolecolare, ricerche sul Genoma, simulazione dell'ecosistema geofisico terrestre, sismologia, scienze dei materiali, medicina, *image processing* e creazione di contenuti digitali (animazioni e scene di film create a computer), servizi finanziari, simulazione di

Il supercomputer più potente del mondo e il più potente d'Italia

Data Center	Los Alamos National Laboratory (USA)	CINECA (Italia)
Posizione Top 500	1	177
Tipo di Sistema	IBM Cluster	IBM Cluster
Modello	RoadRunner	Cluster BladeCenter HS21, Xeon dual core IBM BCX/5120
Prestazioni	Sustained: 1,0 PetaFlops (1000 TFlops) Picco: 1,71 PetaFlops	Sustained: 26,6 TeraFlops Picco: 61 TeraFlops
Memoria principale	103,6 TB	10,24 TB
Anno Installazione	2008	2008
Sistema Operativo	Linux	Linux Red Hat
Connessione	Infiniband	Infiniband
Processori	12.960 CPU IBM PowerXCell 8i, 6.480 processori AMD Opteron dual-core	2.560 Opteron Dual Core 2,6 GHz

reti elettriche e di TLC, progettazione/simulazione e di prodotti e processi di ogni tipo: per esempio Procter & Gamble utilizza l'HPC per modellare la produzione delle patatine Pringles e dei pannolini Pampers.

La Hit Parade dei più potenti

Dal 1993, Jack Dongarra dell'Università del Tennessee mantiene la lista dei 500 maggiori supercomputer del mondo e delle rispettive architetture e applicazioni (TOP500, www.top500.org). Le prestazioni sono misurate in base al **Benchmark LINPACK**, ideato dallo stesso Dongarra, che risolve un sistema di equazioni lineari e utilizza come unità di misura i Flops.

L'attuale sistema N. 1 (novembre 2008) è il cluster linux costruito da IBM per il Los Alamos National Laboratory chiamato "Roadrunner," il primo supercomputer che ha superato la barriera del Petaflops (10^{15} Flops), e che assorbe "solo" 2,35 MW per alimentare 296 armadi che occupano 6.000 piedi quadrati (pari a 560 m^2). Il sistema Roadrunner è basato su blade che utilizzano una versione avanzata dei processori della PlayStation 3 Sony, e relega al secondo posto il precedente primatista IBM BlueGene/L del Lawrence Livermore National Laboratory (478,2 TeraFlops).

Nella lista Top 500 compaiono anche 10 supercomputer italiani. Il più potente supercomputer italiano (posizione 177) è il cluster IBM-BCX/5120 installato nel 2008 al CINECA di Bologna, dotato di 2560 processori AMD dual-core (5120 core in totale) con interconnessione Infiniband, che viene usato principalmente per applicazioni massicciamente parallele e progetti industriali di punta.

Sala macchine CINECA

Le sfide per il futuro

Dopo avere raggiunto l'obiettivo dei **PetaFlops**, il nuovo obiettivo sono gli **ExaFlops**, un miliardo di miliardi di operazioni in virgola mobile per secondo (10^{18} Flops), il cui raggiungimento è previsto intorno all'anno 2018.

Questo comporta la necessità di risolvere nuovi problemi, poiché lo sviluppo dei microprocessori in generale e dei supercomputer in particolare è arrivato a un momento di svolta, dovuto al raggiungimento di alcune barriere fisiche.

La barriera più importante è quella della potenza assorbita (Power wall) e del calore prodotto di conseguenza, che dipende dal numero e dalla densità di transistor e dalla velocità di clock (i supercomputer con potenze di calcolo di PetaFlops/ExaFlops assorbiranno fino a 2 o 3 GWatt).

Mentre tradizionalmente i tempi di elaborazione erano maggiori dei tempi di trasferimento con le memorie, ora, grazie agli enormi progressi degli acceleratori, si ha il fenomeno contrario (Memory wall). È quindi necessario velocizzare i sistemi di trasferimento tra memoria e CPU.

Con le geometrie di 65 nanometri o più dense, aumenta anche la "potenza statica" assorbita in condizioni di riposo dalle correnti parassite, che sviluppa una quantità di calore proporzionale al quadrato della frequenza di clock, e aumentano anche gli errori dovuti ai disturbi e i guasti.

Architetture di High Performance Computing

Supercomputer Sistema unico, con varie CPU e acceleratori (unità vettoriali, GPU, etc)	Costruito da un unico sistema multiprocessore Shared memory processing (SMP) Adatto per applicazioni poco parallelizzabili Processori ottimizzati per il supercalcolo
Cluster Un gruppo di elaboratori collegati da una rete dedicata che coordina le loro azioni per fornire servizi scalabili e alta affidabilità.	È costituito da più di un computer Distributed memory processing Elevato parallelismo
Grid Rete locale o geografica di elaboratori, coordinati da un programma supervisore comune.	"Internet is the computer" Nessuna limitazione geografica. I collegamenti tra i nodi possono essere più lenti che tra i nodi di un cluster
Cloud "Nuvola" di risorse di elaborazione anche eterogenee, che fornisce servizi SaaS	L'architettura è trasparente. Si vedono solo i servizi erogati dalla "nuvola".

Il problema viene attenuato utilizzando microprocessori *multicore* con una frequenza di clock più bassa, raffreddamento di precisione a fluido e materiali semiconduttori in grado di lavorare a temperature più alte senza diminuire troppo l'affidabilità (secondo l'equazione di Arrhenius, un aumento di temperatura di 10° C comporta un raddoppio della frequenza dei guasti).

È anche necessario migliorare le funzione di autodiagnosi e la robustezza (*fault tolerance*) di hardware, software e architetture.

Inoltre occorre produrre software più elastico, adattivo e asincrono e fare evolvere le librerie numeriche, continuando il cammino iniziato con le librerie vettoriali **Linpack** degli anni '70, e proseguito con le librerie a memoria condivisa **Lapack** degli anni '80 e a memoria distribuita **ScaLapack** degli anni '90, fino alla nuova generazione **Plasma** per i chip multicore, per sfruttare le nuove architetture distribuite su scala geografica Grid e Cloud Computing.

Grid computing

Una Grid è una federazione di sistemi di elaborazione dotata di un middleware di supervisione delle risorse di calcolo, di memoria e degli utenti della grid, che ha la funzione di accoppiare le risorse richieste e quelle disponibili per garantire la distribuzione ottimale dei carichi in funzione dello stato dell'intera grid (*match-making*). Attualmente, la più importante grid europea è EGEE del CERN di Ginevra.

Accanto alle grandi Grid nazionali e internazionali, esistono molteplici implementazioni di sistemi distribuiti con architettura Grid su scala locale o metropolitana, definiti Local Area Grid (LAG) e Metropolitan Area Grid (MAG), che si avvicinano al concetto di Intranet e forniscono un'infrastruttura che può essere usata per il calcolo distribuito in ambito aziendale.

L'organismo che si occupa dello sviluppo degli standard relativi alle grid è GGF (Global Grid Forum – www.ggf.org), che ha creato il modello di riferimento OGSA (Open Grid Services Architeture). Il software open source di Grid Computing più utilizzato è BOINC (Berkeley Open Infrastructure for Network Computing), sviluppato dall'Università di Berkeley.

Cloud Computing

La tecnologia Cloud Computing si affianca e si contrappone al Grid come modalità di utilizzazione in modo trasparente e condiviso di risorse di calcolo distribuite geograficamente.

Il Cloud Computing va ad affiancare, senza sostituirla, la tecnologia Grid, con cui ha in comune molti concetti di base, benché con presupposti e obiettivi diversi.

Infatti, mentre il Grid è più orientato al mondo della ricerca, il Cloud ha un approccio più commerciale e punta a rendere fruibili in modo semplice e diretto le risorse di calcolo a coloro che ne fanno richiesta, pagando in base all'utilizzo (*pay-per-use*). Il Cloud si basa sulla virtualizzazione delle risorse di calcolo per fornire all'utilizzatore finale un servizio che può essere usato senza conoscere i dettagli tecnici dei sistemi. Tra i più significativi utilizzatori del Cloud Computing compaiono Google Apps Engine e New York Times.