Confronto tra Opteron single-core e dual-core

Marino Marrocu,
Marco Talice,
Riccardo Triunfo,
Fabio Maggio,
Carlo Podda,
Alan Scheinine

CRS4
Centro di Ricerca, Sviluppo e Studi Superiori in Sardegna
Polaris, Edificio 1
Loc. Pixina Manna
09010 Pula (Cagliari), Italia

September 2005

Contents

1 Misure di Prestazione

L'utilizzo di diverse architetture può segnalare se la comunicazione o l'accesso alla memoria principale sarebbero un collo della bottiglia. Un esempio è la CPU "dual-core" in cui la prestazione di calcolo è raddoppiata mentre la largezza di banda alla memoria rimane invariata; e quindi, un programma in cui l'accesso alla memoria principale è un collo della bottiglia non dimostrerebbe un aumento di prestazione quando una CPU dual-core sostituisce una CPU single-core.

Il CRS4 ha un computer in cui la scheda madre ospita due CPU di tipo Opteron dual-core. Questo tipo di CPU contiene due processori nel stesso "chip". Anche il cache a primo livello (L1) e secondo livello (L2) sono replicati, cioè ogni processore (core) ha il suo proprio cache L1 e L2. Però l'accesso alla memoria sulla scheda madre rimane uguale a single-core e quindi i due processori sullo stesso chip debbono condividere il canale alla memoria. La macchina dual-core usata per le prove (benchmarking) ha due chip per un totale di quattro processori e due canale alla memoria principale.

Inoltre, il CRS4 dispone di un quad-Opteron single-core, cioè una macchina con quattro CPU sulla stessa scheda madre, in cui le CPU condividono la memoria principale. Ogni CPU ha un canale verso un banco di memoria con la conseguenza di avere quattro processori (nel senso di "core") e quattro canali alla memoria principale.

Una terza configurazione consiste di un cluster di macchine con scheda madre a due CPU single-core, in particolare il cluster per la meteorologia. Con due processi su un nodo del cluster e due processi su un altro nodo, si aspetta di vedere una prestazione simile alla macchina quad-Opteron quando la comunicazione tra nodi non è un collo della bottiglia. Il cluster per la meterologia ha due tipi di connessione per la comunicazione tra processi MPI: Gigabit Ethernet e Myrinet. Myrinet fornisce due Gigabit/sec ed una latenza minore in confronto a Gigabit Ethernet. Quindi il confronto sul tempo di esecuzione tra l'utilizzo di Gigabit Ethernet e l'utilizzo di Myrinet è un altro metodo per misurare se la prestazione è vincolata dalla comunicazione.

Le specifiche dei computer usati per il benchmarking sono le seguenti:

Computer dual-core. Scheda madre con due chip CPU, ogni chip di tipo dual-core per un totale di quattro processori core. Ogni processore ha il suo proprio cache però la largezza di banda alla memoria principale è uguale a single-core, cioè, il numero di banchi di memoria (two-way interleaved) è uguale al numero di chip. La CPU è dual-core AMD Opteron modello 875 con frequenza di 2,2 GHz. (Il modello 275 sarebbe normale per questa scheda però soltanto il 875 era disponibile quando

questo computer è stato assemblato.) Il sistema operativo è CentOS 4.1 con due kernel di Linux, versione 2.6.9-11.ELsmp e versione 2.6.12.2. Le prime prove sono state eseguite con il kernel della distribuzione: 2.6.9-11.ELsmp. In una conversazione nel mailing list Beowulf B è stato indicato che il kernel 2.6.12 ha dei migliormenti per dual-core e non-uniform memory access (NUMA). Un secondo gruppo di prove sono state eseguite con il recent kernel 2.6.12.2.

Computer quad-Opteron di nome bithia. Scheda madre con quattro chip CPU, ogni chip di tipo single-core per un totale di quattro processori core. Questa scheda ha quattro banchi di memoria. La CPU è modello Opteron 848 con una frequenza di 2,2 GHz. Il sistema operativo è una distribuzione costruito dal CRS4, basato su Rocks Clusters, con Linux kernel versione 2.6.7.

Cluster di nome scivu. Ogni nodo del cluster ha una scheda madre con due chip CPU, ogni chip di tipo single-core. Una prova con quattro processi userebbe due nodi collegati con Gigabit Ethernet oppure Myrinet. La CPU di modello AMD Opteron 244 su questi nodi hanno una frequenza di 1,8 GHz. Il sistema operativo è una distribuzione costruito dal CRS4, basato su Rocks Clusters, con Linux kernel versione 2.6.10.

Specifica della memoria. Per tutte le schede madre coinvolte nelle prove, la memoria principale è uguale, DDR a 400 MHz con accesso "interleaved" tra due moduli.

1.1 Programma Bolam

1.1.1 Dettagli sul Benchmarking

Per le prove di benchmarking con il programma "Bolam" di meteorologia, quattro processi di MPI (versione MPICH I) sono stati avviati. La tabella 1.1 dimostra i tempi di esecuzione. Le opzioni di compilazione erano "`-O2 -fast -Minline`". Le opzioni descritte in Sec. 1.2 fanno un insieme migliore, la scelta per Bolam corresponde all'insieme usato durante la parallelizzazione del programma quando il compilatore fu versione 5.1 della PGI. Il compilatore impegnato per il benchmarking era pgf90 version 6.0-1 della Portland Group. I file con i dati erano messi in una cartella locale per la macchina dual-core e per bithia. Ogni "run" è stato eseguito due volte perchè dopo la prima volta i dati dei file rimangono in memoria, e questo può avere un impatto sul tempo di esecuzione.

Computer	Linux kernel	tempo CPU (sec.)	tempo reale (sec.)
dual-core, 1^o run	2.6.9-11.ELsmp	133	137
dual-core, 2^o run	2.6.9-11.ELsmp	122	128
dual-core, 1^o run	2.6.12.2	77	81
dual-core, 2^o run	2.6.12.2	64	77
bithia, 1^o run	2.6.7	67	79
bithia, 2^o run	2.6.7	69	79
due scivu Gb Ethernet, 1^o run	2.6.10	69	84
due scivu Gb Ethernet, 2^o run	2.6.10	70	82
due scivu Myrinet, 1^o run	2.6.10	68	71
due scivu Myrinet, 2^o run	2.6.10	65	66

Table 1.1: Ogni run usava quattro processi con la comunicazione di tipo message-passing di MPI. Le prime due righe corresponde a una scheda con due chip di tipo dual-core. La macchina bitha contiene una scheda madra con quattro chip single-core. Per il cluster scivu, sono impegnato due nodi sia con la connessione Gigabit Ethernet che con la connessione Myrinet.

1.1.2 Discussione

Quando Bolam è stato fatto girare in un cluster con i node con scheda madre con le CPU Athlon a 32 bit (due CPU per nodo), si è verificato che il programma era più efficiente quando ogni nodo aveva soltanto un processo. Il tempo di esecuzione si aumentava quando un nodo aveva due processi (un processo per ogni CPU). Questo effetto indicava che il programma Bolam ha bisogno di un ampia largezza di banda alla memoria principale. La necessità di avere soltanto un processo per nodo non si verifica sul cluster scivu in cui i nodi hanno il processore Opteron single-core con un largezza di banda a memoria maggiore alla Athlon. Però la largezza di banda a memoria diventa un problema per i chip dual-core se il kernel di Linux non è aggiornato.

Il fatto che il tempo di esecuzione sul cluster scivu (con l'utilizzo di Myrinet) con le CPU a 1,8 GHz è simile al tempo di esecuzione sulla macchina bithia con le CPU a 2,2 GHz è un'altra indicazione che la prestazione è vincolata dalla velocità di accesso di memoria, sempre a 400 MHz DDR interleaved. Però non si può escludere che l'implementazione di message passing sulla bithia non era al massimo dell'efficienza.

1.2 Programma Non-Idrostatico

1.2.1 Dettagli sul Benchmarking

Computer	Linux kernel	tempo CPU (sec.)	tempo reale (sec.)
dual-core, 1^o run	2.6.9-11.ELsmp	536	574
dual-core, 2^o run	2.6.9-11.ELsmp	508	544
dual-core, 1^o run	2.6.12.2	511	585
dual-core, 2^o run	2.6.12.2	430	476
bithia, 1^o run	2.6.7	396	457
bithia, 2^o run	2.6.7	390	453
due scivu Gb Ethernet, 1^o run	2.6.10	451	535
due scivu Gb Ethernet, 2^o run	2.6.10	462	528
due scivu Myrinet, 1^o run	2.6.10	422	426
due scivu Myrinet, 2^o run	2.6.10	422	428

Table 1.2: Ogni run usava quattro processi con la comunicazione di tipo message-passing di MPI. Le prime due righe corresponde a una scheda con due chip di tipo dual-core. La macchina bitha contiene una scheda madra con quattro chip single-core. Per il cluster scivu, sono impegnato due nodi sia con la connessione Gigabit Ethernet che con la connessione Myrinet.

Per le prove di benchmarking con il programma "Non-Idrostatico" di meteorologia, quattro processi di MPI sono stati avviati. La tabella 1.2 dimostra i tempi di esecuzione. Le opzioni di compilazione erano "`-O3 -fast -Mvect=sse -Mvect=prefetch -Mcache_align -Mipa -Minline`". Il compilatore impegnato era pgf90 version 6.0-1 della Portland Group. I file con i dati erano messi in una cartella locale per la macchina dual-core e per bithia. Ogni "run" è stato eseguito due volte perchè dopo la prima volta i dati dei file rimangono in memoria, e questo può avere un impatto sul tempo di esecuzione.

1.2.2 Discussione

Il confronto tra la macchina dual-core e la macchina quad-Opteron (bithia) indica che l'accesso alla memoria principale è un vincolo però non tanto grave quanto il caso di Bolam. Facendo confronto tra Gb Ethernet e Myrinet, vediamo che la comunicazione può essere un collo della bottiglia. Altre prove sul numero di processi MPI ci hanno fornito ulteriori indicazioni che la comunicazione sia vincolante.

Resta notevole il fatto che il tempo reale di esecuzione su un cluster con Myrinet è superiore al tempo reale sul una macchina (bithia) con quattro Opteron sulla stessa scheda, nonostante la frequenza maggiore delle CPU di bithia. Forse altre opzioni di compilazione avrebbero dato una prestazione superiore su bithia. Non abbiamo seguito l'ottimizzazione per bithia perché il costo per CPU di una scheda quad-Opteron è quasi il doppio del costo per CPU di una scheda con due Opteron, e quindi finché questo differenza non va ridotta, non prevediamo l'utilizzo di schede quad-Opteron per questi programmi.

1.3 Programma Fluidodinamica

1.3.1 Dettagli sul Benchmarking

Computer	Linux kernel	tempo CPU (sec.)	tempo reale (sec.)
dual-core, *opt1*	2.6.9-11.ELsmp	5172	5608
dual-core, *opt1*	2.6.12.2	3042	3553
bithia, *opt1*	2.6.7	3689	4347
due nora Gb Ethernet, *opt1*	2.6.7	2695	3681
due scivu Gb Ethernet, *opt1*	2.6.10	3765	4664
due scivu Gb Ethernet, *opt2*	2.6.10	2969	3940
due scivu Myrinet, *opt1*	2.6.10	(n/d)	4251
due scivu Myrinet, *opt2*	2.6.10	(n/d)	2896

Table 1.3: Ogni run usava quattro processi con la comunicazione di tipo message-passing di MPI. Le prime due righe corresponde a una scheda con due chip di tipo dual-core. La macchina bitha contiene una scheda madra con quattro chip single-core. Per il cluster nora, sono impegnato due nodi con la connessione Gigabit Ethernet. Per il cluster scivu, sono impegnato due nodi sia con la connessione Gigabit Ethernet che con la connessione Myrinet. Il simbolo *opt1* significa le opzione di compilazione di "`-O2 -fast`" mentre il simbolo *opt2* significa le opzione di compilazione di "`-O3 -fast -Mvect=sse -Mvect=prefetch -Mcache_align -Mipa -Minline`". Il simbolo "(n/d)" significa che nessun dato è disponibile.

Per le prove di benchmarking con il programma "Fluidodinamica" quattro processi di MPI (versione MPICH I) sono stati avviati. La tabella 1.3 dimostra i tempi di esecuzione. Il compilatore impegnato per il benchmarking era pgf90 version 6.0-1 della Portland Group.

In altro modo di fare confronto tra la prestazione single-core e dual-core è di consid-

erare la prestazione quando o 2 task o 4 task girano su un unico nodo del cluster.

- single-core, 4 tasks / 2 processors, 6187 sec. tempo reale
- dual-core, 4 tasks / 2 processors, 3553 sec. tempo reale
- single-core, 2 tasks / 2 processors, 2979 sec. tempo reale
- dual-core, 2 tasks / 2 processors, 2570 sec. tempo reale

1.3.2 Discussione

Non tutte le prove sono eseguite con le opzioni migliori. Comunque i dati dimonstrano che un nodo con i chip dual-core equivale a due nodi con i chip single-core della stessa frequenza. Inoltre, quando l'ottimizzazione del compilatore è al massimo, vediamo un vantaggio con l'utilizzo della connessione veloce di Myrinet.

1.4 Programma RepeatMasker

RepeatMasker è un programma utilizzato per rintracciare e mascherare le ripetizioni che possono essere presenti, per esempio, nelle sequenze del DNA. È parallelo ma in multithread, non con il Message Passing (MPI) o PVM. La sua esecuzione non può essere distribuita su nodi remoti. Abbiamo fatto girare RepeatMasker su dualcore, su nora e su bithia a 2 threads e a 4 threads.

Computer	Linux Kernel	num threads	tempo reale (hh.mm.ss)
dualcore	2.6.12.2	2	1.43.00
nora	2.6.12-2	2	1.54.00
bithia	2.6.7	2	1.43.00
dualcore	2.6.12-2	4	0.56.04
nora	2.6.12-2	4	1.56.38
bithia	2.6.7	4	0.58.28

Table 1.4: La macchina dualcore ha due cpu Amd 875 dual core, nora ha due cpu Amd 246 mentre bithia monta 4 cpu Amd 848. L'esecuzione di RepeatMasker è, a due e quattro threads, è stata avviata sul cromosoma 20.

La tabella 1.4 mostra che, per RepeatMasker, una cpu 875, dualcore, equivale nella sostanza a due cpu 848, singlecore. Il divario tra la cpu Amd 875 e la Amd 246 è fondalmentalmente dovuto all'incremento di velocita della 875 (2200MHz) sulla 246 (2000MHz).

1.5 Discussione

L'inserimento di bithia nella lista dei test è dovuto solo alla necessità di confrontare le performance di dualcore al variare del numero dei threads con il comportamente di una quad-cpu vera e propria. I tempi assoluti non sono da confrontare poichè bithia utilizza un kernel non recente e questo senz'altro la penalizza in maniera che non è al momento possibile quantificare. Detto questo la dualcore *scala* molto al crescere del numero dei threads, come una vera e propria macchina a quattro cpu.

1.6 Programma ELASEM (Parallel Linear Elasticity with Spectral Elements)

ELASEM è un solutore, parallelo, per equazioni dell'acustica in 2D. Il programma utilizza MPICH per la comunicazione. I test con ELASEM sono stati fatti con 2 e 4 task paralleli su una nora, sulla dualcore e con 4 tasks su altre due nora. Anche nel test con

Computer	Linux Kernel	tasks	tempo reale (sec)
dualcore	2.6.12.2	2	52,69
nora	2.6.12-2	2	63,78
dualcore	2.6.12-2	4	31,11
nora	2.6.12-2	4	61,30
2 nora eth	2.6.7	4	38,80

Table 1.5: La macchina dualcore ha due cpu Amd 875 dual core, le nora hanno due cpu Amd 246. La comunicazione tra i tasks avviene tramite MPICH. Le due nora comunicano tramite Giga ethernet. ELASEM è stato lanciato a due e quattro task MPI.

il solutore ELASEM, i cui risultati vediamo nella tabella 1.5, viene mostrato che una cpu 875, dualcore, equivale nella sostanza a due cpu 246, singlecore.