ISTITUTO NAZIONALE DI FISICA NUCLEARE

Sezione di Padova

INFN/CCR-07/11
November 30, 2007

CCR-18/2007/P
Settembre 2007
Versione 1.4

COSTI E PRESTAZIONI DEI WORKER NODE PER IL CALCOLO LHC

Michele Michelotto[1)]

[1)]*INFN-Sezione di Padova, Via F. Marzolo, 8, I-35131 Padova, Italy*

Abstract

Questo documento descrive il panorama dei processori disponibili sul mercato per nodi di calcolo (Worker Node) nei prossimi sei – dodici mesi.

La valutazione delle prestazioni dei processori in questo momento di transizione tra SPEC int 2000 e SPEC int 2006 è molto delicato perché non è chiaro quale sarà la scelta futura come benchmark di riferimento.

Spero in un rapido passaggio a SPEC int 2006 rate misurato con compilatore gcc.

Cercherò anche di prevedere i prezzi per acquisti tipici da Tier2 (10-30 box).

PACS.: 89.20-Ff (Computer Science and Technology)

Published by ***SIS–Pubblicazioni***
Laboratori Nazionali di Frascati

1 INTRODUZIONE

Lo scopo di questo documento è quello di presentare un panorama dei processori disponibili sul mercato per nodi di calcolo (Worker Node) nei prossimi sei – dodici mesi. Cercherò anche di prevedere i prezzi per acquisti tipici da Tier2 (10-30 box).

Vedremo le attuali tecnologie dei semiconduttori, l'influenza della legge di Moore sul clock, dimensioni delle cache e consumi energetici. Il tipico worker node sarà ancora dual processor con processori dual core o quad core e 2GB per ogni core. Le configurazioni a blade risultano ancora leggermente più costose, ma più convenienti in termini di costi totali di esercizio a causa di consumi ridotti del 20 – 30% rispetto ad analoghe configurazioni pizza-box, ovviamente questo è vero solo se si acquistano un numero di blade tali da ammortizzare il costo del crate.

2 UNITÀ DI MISURA

L'unità di misura delle prestazioni in questo momento è un aspetto molto critico. Le richieste degli esperimenti e le potenze promesse dai centri di calcolo Tier1/Tier2 sono in termini di Specint 2000, cioè espressi in termini di CPU Int 2000 della suite SPEC.

Da qualche mese la suite CPU 2000 è stata sostituita da SPEC con la suite CPU2006 per obsolescenza tecnologica. In pratica, gli avanzamenti tecnologici delle CPU, delle loro cache e delle memorie RAM dei sistemi rendono inaffidabili le valutazioni di performance in termini di CPU 2000 int.

Per esempio i processori Intel dual core 51xx e i processori AMD dual core 22xx danno prestazioni molto simili quando si fanno girare programmi applicativi HEP CPU bound, mentre la prima CPU ha un rating doppio della seconda, utilizzando gli SPEC INT 2000 pubblicati dal sito spec.org.

Questa enorme differenza è dovuta soprattutto all'uso del compilatore gcc al posto di compilatori commerciali (la tipica configurazione di misura consiste di sistema operativo windows e compilatore intel): Se confrontiamo per esempio Specint2000 calcolato sui vecchi processori Xeon e sui nuovi Woodcrest e Clovertown vediamo una perdita fino a 55-65% del gcc (compilato con bassa ottimizzazione) rispetto a icc.

Specint2000/GHz	gcc –O2 –fPIC -pthread	icc –fast & pgo	Rapporto
Nocona 32bit	250.36	422.86	59%
Dempsey 64 bit	305.56	472.19	65%
Woodcrest 32 bit	530.76	969.24	55%
Woodcrest 64 bit	637.28	1005.63	63%
Clovertown 32 bit	637.56	990.61	64%
Clovertown 64 bit	623.75	981.67	64%

Ma si osservano grandi variazioni anche con gcc variando

- La versione del compilatore ***gcc3*** vs ***gcc4.***

- Codice a 64 bit vs codice a 32 bit (invece non dipende dal fatto che il sistema operativo sia a 32bit o a 64 bit, per codice a 32bit). Qui non ho molti dati ma per esempio ottengo su un Woodcrest 1390 rispetto a 1556 (81%) o su un AMD 275 750 rispetto a 1045 (71%), rispettivamente a 32bit e 64 bit.
- Gli switch di compilazione: compilando *gcc –O2 –fPIC -pthread* causava un grande drop di prestazioni, soprattutto su processori intel a 32bit rispetto a *gcc –O3*. Questo perché il CERN cerca di usare gli stessi switch di compilazione della maggior parte dei programmi LHC. Di nuovo non ho molti dati ma per esempio un AMD275 perdono il 35% delle prestazioni con il CERN tuning rispetto a un "normale" "gcc –O3" mentre gli intel perdono il 20-25%. Quindi usare questo tuning è particolarmente penalizzante per gli AMD.

Per tutti questi motivi il CERN da qualche tempo richiede che i partecipanti ai tender per l'acquisto dei Worker Nodes forniscano gli specint misurati con gcc *–O2 –fPIC –pthread*.

In realtà ritengo necessario passare al più presto all'uso di Spec CPU int 2006 perché è meno sensibile di CPU Int 2000 alle dimensioni delle cache di secondo livello ed è pensato per occupare circa 1GB per ogni core contro i 200 MB/core di CPU int 2000.

In pratica l'uso di CPU2000 int con i dati pubblicati dal sito della SPEC porta a risultati inaffidabili.

Questo vuol dire che quando si andranno a considerare i prezzi dovremo prestare attenzione al prezzo in termini di Euro/SI2000 sia per confronti con gli acquisti precedenti, sia perché questa è ancora adesso l'unità di misura ufficiale (CERN, INFN, WLCG) usata negli accordi tra esperimenti e agenzie.

Ma se vogliamo ottenere le migliori macchine dobbiamo guardare soprattutto al prezzo in termini di Euro/SI2006.

Per dare un termine di paragone il rapporto tra SI2000 e SI2006 è compreso tra 165 e 175 per macchine Intel di ultima generazione e tra 140 - 145 per macchine AMD di ultima generazione (22xx).

Il discorso in realtà potrebbe essere ancora più complicato dal momento che dovremmo tenere conto anche della metrica CPU INT RATE con cui si misurano le prestazioni di throughput della macchina invece della semplificazione di moltiplicare le prestazioni single threaded CPU INT per il numero di core. Questo non causa grandi problemi dal momento che per job CPU bound le prestazioni crescono linearmente almeno fino a macchine con doppio quad core.

In questo momento la situazione è piuttosto confusa. Il CERN continua a fare gare basandosi su SPEC Int 2000 misurato con gcc. GridKA che gestisce il Tier1 Tedesco sembra essere orientato verso Specint 2006. Il mio consiglio è di usare SPEC Int 2006 rate misurato con gcc o megli SPEC Int 2006 rate.

Per il prezzo invece considero il prezzo IVA Compresa in Euro.

3 IL NODO DI CALCOLO

Il nodo di calcolo più conveniente è il box 1U biprocessore con un alimentatore e un piccolo disco ATA (o due dischi ATA in mirror).

Ogni processore può avere due o quattro cores, permettendo quindi di processare rispettivamente quattro o otto job in parallelo.

L'alternativa è la configurazione blade in cui i box vengono infilati in verticale in

cestelli (crate) dotati di alimentatori ridondati ad alta efficienza. Il vantaggio della configurazione blade sta nel fatto che tutti i nodi continuano a funzionare se uno degli alimentatori si rompe (Se un nodo con 8 core si rompe invece rimango con 8 code inutilizzabili).

Inoltre il blade può avere un unico accesso fisico locale (risparmio di costosi switch KVM) e un unico accesso IPMI. Inoltre spesso sono dotati di switch Ethernet integrati.

Lo svantaggio sta soprattutto nel fatto che si devono acquistare un numero minimo di blade per ammortizzare l'acquisto dell'infrastruttura (alimentatori, crate, switch) e nel fatto che nel futuro si rimane vincolati ad un certo modello di blade su cui si è già investito.

Dal punto di vista dei costi le configurazioni blade costano 10 – 20% più delle configurazioni 1U a cestelli pieni. Tuttavia il costo totale di esercizio dei blade viene pubblicizzato (IBM, HP, Supermicro) come più conveniente, dal momento che mentre gli alimentatori delle macchine 1U hanno efficienza energetica vicino al 75%, quella degli alimentatori dei blade sono superiori al 90%.

Esempio: Se ho un carico di 500 Watt un box 1U con efficienza del 75% ha bisogno d 667 Watt di ingresso mentre un blade richiederebbe solo 538 Watt. Un risparmio di 129 Watt che in un anno a pieno carico corrispondono a 1130 kWh. Se prendiamo come costo del kWh 0.15 Euro abbiamo un risparmio di 169.50 Euro/anno.

Sui tre anni di vita di una macchina, sull'acquisto di un cestello di 10 nodi avremmo un risparmio di 5085 Euro. Senza contare il risparmio dovuto al minor consumo di raffreddamento e UPS.

Infatti eventuali risparmi energetici non incidono solo sulla bolletta elettrica per i nodi di calcolo nei tre/cinque anni di vita delle macchine ma anche nelle minori richieste di impianti tecnologici infrastrutturali (UPS, Gruppi Elettrogeni, Condizionatori) e di conseguenza nel minor consumo elettrico degli impianti stessi.

4 TWIN SERVER

Un'altra possibilità è data dalle configurazioni twin. In un case 1U vengono inseriti due motherboard simili a quelle dei blade, molto strette e lunghe, disposte affiancate. In pratica è come avere due blade in orizzontale. Il vantaggio rispetto ai blade sta nel risparmiare tutti i costi di infrastruttura del cestello. Tuttavia abbiamo in pratica due server stipati nello stesso blade con un unico alimentatore.

Chi scegliesse configurazioni di questo tipo deve prestare molta attenzione alla dissipazione termica. Infatti queste macchine hanno alimentatori da 900 Watt con consumi prossimi a 500 Watt. Un rack pieno di queste macchine deve essere servito da condizionatori che possano rimuovere oltre 20 kW.

5 I PROCESSORI ATTUALI

I processori disponibili in questo momento sono i seguenti.

Woodcrest: la serie dual core della **Intel 51xx**, clock da 2000 MHz (5130) a 3000 MHz (5160). Ogni core ha 32+32KB di cache di primo livello e 2 MB di cache di secondo livello. Tecnologia a 65 nm. Consumi contenuti per quanto riguarda il processore ma richiede Memory controller e Memoria FBDIMM che consumano parecchio.

Esempio: due processori 2x65 W + memory controller 32 W + chipset 12 W +

FBDIMM 83 W = **Totale server 257 Watt**

Clovertown: la serie quad core della **Intel 53xx**, clock da 2000 a 2666 MHz. Stessa cache del Woodcrest. In pratica è come se ci fossero due chip Woodcrest affiancati nello stesso chip enclosure e quindi in un socket.

Opteron Socket F. La serie dual core di AMD. Hanno cache di primo livello più grandi (64KB+64KB ma solo 1MB per core di cache di secondo livello. La tecnologia è ancora a 90nm. Clock da 1800 (2210) a 3000 MHz (2222SE). Consuma più della controparte intel ma poi non necessita di Memory Controller e usa memoria DDR2 che consumano meno.

Esempio di consumo: due processori 2 x 85W + 2 banchi di memoria 2x18W + chipset 16W = **Totale server 242 Watt**.

6 IL PROSSIMO FUTURO

Sono stati annunciati recentemente i processori quad core di AMD. Dovrebbero essere disponibili in volumi da Settembre con il nome in codice “Barcelona” o K10. Questi sono dei veri processori quad core in unico chip con un’interessante architettura con cache di primo di livello di 64+64KB, una cache di 512 MB LVL2 per core e 2 MB di cache LVL3 condivisa dai 4 cores.

Quindi una cache di secondo livello più piccola ma con la possibilità di un altro livello di cache on-chip condivisa tra i cores.

Clock iniziali attorno ai 1.6 GHz - 2GHz per poi salire verso i 3GHz con gli stepping successivi. Consumi tra 68 e 120 Watt.

A prima vista, il vantaggio del quad core AMD sta nella possibilità di comunicare direttamente tra i core attraverso il bus proprietario di interconnessione, mentre il quad core Intel dovrebbe passare attraverso il Front End Bus.

Ci saranno anche miglioramenti nelle prestazioni Floating Point.

D’altra parte Intel ha clock più alti e cache LVL2 maggiori, soprattutto per la possibilità di sfruttare appieno la tecnologia a 65 nm.

7 FINE 2007 – INIZIO 2008

Peynrin: A fine 2007, inizio 2008 sono previsti anche le versioni a 45 nm di questi chip. La riduzione delle dimensioni portano a densità più alte e ad una riduzione delle tensioni sui gate. Di conseguenza i nuovi chip potranno avere da una parte cache on-chip più grandi, e d’altra parte la possibilità di ottimizzare i clock. La scelta è tra clock più veloci a consumi costanti o clock uguali a quelli attuali con consumi più bassi e possibilità di aumentare il numero dei cores a consumi costanti.

Questi processori permetteranno ad Intel di avere dei veri quad core in un singolo chip. Avranno consumi compresi tra 50 e 120 Watt per i quad-core e tra 40 e 80W per i dual core. Il clock dovrebbe essere attorno ai 3.33 GHz (già ora ci sono chip intel desktop a 3.8 GHz).

Si prevedono cache fino a 6MB per chip nei dual core e 12 MB nei quad core (1.5 per core).

Nel Q2 2008 AMD dovrebbe rispondere con il processore con il nome in codice **Shangai** con 6 MB di cache LVL3 (contro i 2MB del Barcelona) costruito in tecnologia 45 nm.

8 OLTRE IL 2008

La famiglia successiva di intel ha per ora il nome di codice **Nehalem**: Processori fino a 8 core e con il ritorno della possibilità di avere due thread per ogni core. Il multi-threading dovrebbe essere migliore di quello proposto in passato chiamato HyperThreading e poi abbandonato. Questo processore potrebbe avere il memory controller integrato come gli attuali AMD.

Naturalmente andrà verificata la possibilità di scalare da parte del codice HEP che attualmente viene fatto girare come se le macchine fossero single threaded, un job per ogni core. Una macchina con dual 8-core avrebbe 16 job in parallelo o 32 utilizzando il multi-threading.

9 MEMORIA

Tutte le macchine analizzate con quattro core avevano anche quattro Gigabyte di memoria. La tendenza finora è stata quella di avere un Gigabyte per ogni core. Tuttavia l'esperimento Alice ha da tempo necessità di avere 2 GB per core. Ovviamente questo implica un costo per SPEC INT maggiore dal momento che gli SPEC INT sono basati in pratica unicamente sul processore.

Il passaggio del codice degli esperimenti da 32bit a 64 bit dovrebbe permettere un aumento di prestazioni tra il 15% e il 40% ma anche un aumento delle necessità di memoria che non riesco a quantificare (tra 10% e 100%).

Per questo motivo consiglio di acquistare tutte le macchine con 2 GB per core a meno che non sia chiaro che al proprio esperimento 1 GB per core sia più che sufficiente anche nei prossimi tre anni.

10 INDAGINE

Ho chiesto in primavera 2007 a tutte le sedi Tier2 e a Babar i costi e le prestazioni delle macchine relative agli ultimi acquisti. Ho avuto risposte da LNL, Catania, Roma1, Torino, Pisa, Napoli e Babar-Padova.

I prezzi pagati sono circa di 0.45-0.55 Euro/SI2000 per i processori AMD e 0.34 – 0.44 Euro/SI2000 per i 51xx (sappiamo però che gli SI2000 dei 51xx sono inflazionati). In realtà il Tier1 addirittura è sceso a 0.26 con dei 5140.

Ho contattato alcuni venditori di hardware, informandomi su configurazioni 1U con doppio dual core o doppio quad-core e 2GB per core. Per Settembre le previsioni sono di

AMD 22xx 0.45 – 0.48 Euro/SI2000
Intel 51xx 0.32 – 0.34 Euro/SI2000
Intel 53xx 0.22 – 0.25 Euro/SI2000

Rispetto ai Woodcrest risulterebbero più conveniente i Clovertown (ovviamente ho il

doppio di core a parità di tutto il resto o quasi) con prezzo/prestazioni simili a quello ottenuto dal Tier1 a Ottobre 2006.

Se rifacciamo le stesse misure per i più affidabili Specint 2006 abbiamo

AMD 22xx: 59 – 67 Euro/SI2006
Intel 51xx: 57 – 62 Euro/SI2006
Intel 53xx: 38 – 43 Euro/SI2006

Non ho avuto previsioni per i quad-core AMD che immagino saranno più convenienti dei dual core. L'uscita del Barcelona però potrebbe causare una diminuzione del prezzo dei processori Intel e anche il mercato delle memorie potrebbe avere delle oscillazioni sensibili.

Da notare che in passato il Tier1 ha pagato 1.31 Euro/SI2000 nel 2003, 1.29 nel 2004, 0.84 nel 2005 e 0.26 nel 2006.

Prezzo Prestazioni – **Specint pubblicati da Spec.org**

Processore	Euro Lordo	Euro/SI2000	**Euro/SI2006**	cores
AMD Socket F 22xx -8GB	2100-3320	0.40 – 0.45	**55 – 66**	4
Intel Woodcrest 51xx 8GB	2800 - 4050	0.30 – 0.33	**55 – 60**	4
Intel Clovertown 53xx 16GB	3300 - 4600	0.19 – 0.20	**32 – 35**	8
AMD Barcelona 16GB	?? -??	?? - ??	**?? - ??**	8
Twin Clovertown 53xx 32GB	6200-8300	0.18 – 0.19	**31 -32**	16

Secondo questa tabella sembrerebbe che le macchine con processori dual core intel e amd si equivalgano in termini di prezzo prestazioni e che le migliori macchine siano quelle con i processori quad-core intel (in attesa di quelli AMD)

Queste stime per fine 2007 sono abbastanza in linea con le previsioni di B.Panzer del CERN che stimava 0.46 Euro/SI2000 nel 2007 e 0.31 Euro/SI2000 nel 2008, se si tiene conto della sovrastima degli SI2000 nella famiglia dei cores Intel.

Nota bene, ricalcolando la tabella usando i dati SPEC misurati con il tuning CERN:

Prezzo Prestazioni – **Specint misurati con compilatore gcc**

Processore	Euro Lordo	Euro/SI2000	**Euro/SI2006**	Cores
AMD Socket F 22xx -8GB	2400-3600	0.55	**75**	4
Intel Woodcrest 51xx 8GB	2800 - 4100	0.58 – 0.65	**104**	4
Intel Clovertown 53xx 16GB	3800 - 5600	0.39	**65**	8
AMD Barcelona 16GB	?? -??	?? - ??	**?? - ??**	8
Twin Clovertown 53xx 32GB	6200-8300	0.36	**61**	16

In questo secondo caso i processori AMD 22xx risultano più convenienti dei dual core Intel e avvicinano il rapporto prezzo prestazioni dei quad-core Intel.

11 CONCLUSIONI

Il mercato dei nodi di calcolo continua la discesa dei prezzi in termini di Euro/SI2000. Si raccomanda di usare come riferimento "**CPU INT 2006"**.

In questo momento risultano più conveniente le macchine con processore Clovertown rispetto ai Woodcrest e agli Opteron dual core.

Tuttavia non sono ancora disponibile i prezzi dei processori quad-core da AMD che dovrebbero essere sul mercato nell'ultimo trimestre del 2007.

Per evitare ogni ambiguità, i diversi esperimenti dovrebbero misurare le prestazioni reali dei loro codici, prima delle gare, al fine di favorire con il punteggio tecnico l'architettura migliore.