ISTITUTO NAZIONALE DI FISICA NUCLEARE

Sezione di Padova

INFN-12-21/PD
3rd december 2012

SERVER PER NODI DI CALCOLO

Versione 1.0

Settembre 2010

Michele Michelotto[1], Marco Serra[2], Alessandro Brunengo[3]

[1)]INFN-Sezione di Padova, Via F. Marzolo 8, I-35131 Padova, Italy
[2)]INFN-Sezione di Roma, Piazzale Aldo Moro 2, I-00185 Roma, Italy
[3)]INFN-Sezione di Genova, via Dodecaneso 33, I-16146 Genova, Italy

Abstract

Questo documento descrive il panorama dei processori disponibili sul mercato per nodi di calcolo (Worker Node) nel 2010 e previsione per il 2011. Per i nodi di calcolo la metrica principale è HEP-SPEC06[1] e il rapporto Euro/HEP-SPEC06. Il documento è una versione ridotta di un documento redatto dagli stessi autori, su richiesta del Presidente della Commissione Calcolo e Reti come ausilio per i referee delle Commissioni Scientifiche Nazionali con analisi dei costi e stime per gli anni futuri.

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

CCR-43/2012/P

1 INTRODUZIONE

Negli ultimi anni i server per nodi di calcolo sono basati su processori multi-core di Intel e AMD. Gli esperimenti sono in fase di transizione a codice a 64 bit ma l'unità di misura ufficialmente usata per gli esperimenti LHC e per i centri di calcolo da Tier1 ai Tier2 nella GRID INFN è rimasta quella definita da HEP-SPEC06 (di seguito HS06) che è un sottoinsieme di SPEC[2] CPU 2006 compilato a 32 bit.

Il tipico nodo di calcolo per INFN viene consegnato per il montaggio in rack da 19". Il formato meno denso è detto 1U e prevede un server in un "pizza box" alto 1U con due processori.

Per ottimizzare lo spazio rack e avere anche economie di scala ci sono tre tipi di scelta.

- Il formato twin (due server stretti affiancati nello stesso pizza box descritto sopra)
- Il formato doppio twin (quattro server in un box di altezza doppia)
- Il formato blade (un cestello o chassis che contiene diverse blade o lame in verticale). Il costo dello chassis compresi gli alimentatori ridondati e le ventole si aggira sui 3000 – 4000 Euro per cui l'acquisto del blade ha senso solo se il cestello viene riempito quasi completamente. Segnaliamo che esistono anche dei blade con due server ciascuno a bordo.

I processori sono ormai tutti a 64 bit ma permettono di eseguire codice a 32 bit e a 64 bit su sistemi operativi a 64 bit, quasi sempre Scientific Linux. Per cui è prassi comune installare le macchine a 64 bit per semplificare la transizione.

Intel ha una gamma molto vasta di processori da quatto core e sei core. L'ultima serie ha il nome in codice Westemere e code number del tipo 56xx. Quindi è possibile realizzare server con 8 o 12 core totali con clock da 2 a 3 GHz circa. La misura delle prestazioni è particolarmente complicata perché:

- Ogni core può eseguire due thread in parallelo per cui viene visto dal sistema operativo come due cpu logiche abilitando l'Hyperthreading (HT-ON). Tuttavia il throughput del server non raddoppia ma aumenta di circa il 25-30%. Per evitare ambiguità nelle gare probabilmente conviene imporre la misura di HS06 con l'Hyperthreading disabilitato o comunque su un numero di processi concorrenti pari al numero di core fisici.
- Il clock di ogni core varia dinamicamente. Il clock è quello dichiarato quando il processore è a pieno carico, ma se alcuni core sono inattivi questi possono essere spenti o posti in uno stato di consumo estremamente basso mentre i core attivi possono aumentare dinamicamente il loro clock di uno o più incrementi di 133 MHz ciascuno. Quindi per evitare ambiguità le misure vanno fatte con la macchina a pieno carico.

AMD ha presentato nel corso del 2009 i nuovi processori con il nome Magny-Cours e codice di tipo 61xx, ciascuno con 8 o 12 core fisici. Tuttavia i processori di AMD sono ancora in tecnologia a 45 nm contro i 32 nm di Intel ed hanno clock attorno ai 2 GHz.
Nel corso del 2010 sono attesi anche i processori di tipo "Lisbon" con massimo 8 core e codice 4xxx.

Nel corso del 2011 Intel avrà un nuovo tipo di processore, sempre prodotto in tecnologia a 32 nm con un set di istruzioni più esteso e una rinnovata architettura.
Anche AMD avrà una nuova architettura chiamata Bulldozer basata su moduli costituti ciascuno da due core fisici veri e propri che condividono però la parte di FPU. Contribuendo a confondere ulteriormente i concetti di cpu fisica e cpu logica.

2 PROCESSORI INTEL

Il processore di punta di Intel nella fascia DP, dual processor è lo Xeon 56xx noto anche come Westmere. Il precedente processore Xeon 55xx Nehalem è stato l'acquisto preferito nei centri di calcolo HEP nel 2009 ma ormai dovrebbe essere in fase di uscita dal mercato.

La tabella seguente presenta una panoramica dei processori Intel disponibili nel corso del 2010.

I costi dei processori aumentano con le prestazioni in modo non lineare per cui solitamente risultano più convenienti i processori meno potenti. Tuttavia il processore non è l'unico elemento che determina il costo della macchina. Per esempio per la serie 56xx uno dei processori più convenienti è il 5650 che è il più piccolo dei processori con 6 core fisici.

model	Cores/threads	L3 cache	Clock with Turbo Mode Off	TDP
X5680	6/12	12 MB	3.33 GHz	130 W
X5670	6/12	12 MB	2.93 GHz	95 W
X5660	6/12	12 MB	2.80 GHz	95 W
X5650	6/12	12 MB	2.66 GHz	95 W
L5640	6/12	12 MB	2.26 GHz	60 W
X5677	4/8	12 MB	3.46 GHz	130 W
X5667	4/8	12 MB	3.06 GHz	95 W
E5640	4/8	12 MB	2.66 GHz	80 W
E5630	4/8	12 MB	2.53 GHz	80 W
E5620	4/8	12 MB	2.40 GHz	80 W
L5630	4/8	12 MB	2.26 GHz	40 W

Si noti la varietà di combinazioni di clock, potenza dissipata e numero di cores fisici/cpu logiche.

3 PROCESSORI AMD

Analogamente alla tabella precedente per Intel riportiamo qui la tabella equivalente per AMD, e per facilitare il confronto abbiamo lasciato un processore Intel nella prima riga. Si vede bene la differenza dei clock di AMD rispetto a quelli Intel

model	Cores/threads	L3 cache	Clock	ACP
Xeon 5670	6/12	12 MB	2.93 GHz	95 W
6124 HE	8/8	12 MB	1.8 GHz	65 W
6128	8/8	12 MB	1.8 GHz	80 W
6128 HE	8/8	12 MB	2.0 GHz	65 W
6134	8/8	12 MB	2.3 GHz	80 W
6136	12/12	12 MB	2.4 GHz	80 W
6164 HE	12/12	12 MB	1.7 GHz	65 W
6168	12/12	12 MB	1.9 GHz	80 W
6172	12/12	12 MB	2.1 GHz	80 W

6174	12/12	12 MB	2.2 GHz	80 W
6176 SE	12/12	12 MB	2.3 GHz	105 W

Anche qui vediamo diversi clock, numero di core massimi e versioni a basso consumo energetico (HE).

4 MISURE HS06 NELL'INFN

Nel corso del 2010 non ci sono stati molti acquisti all'interno dell'INFN sui processori che abbiano descritto. Abbiamo alcuni dati di HS06 misurati da una macchina (2x Intel Xeon 5650) disponibile a Padova dall'esperimento HEPMARK finanziato da CSN5. Un'altra macchina basata su processori AMD sempre dall'esperimento HEPMARK è in grave ritardo di consegna ma siamo riusciti ad effettuare alcune misure su una macchina resa disponibile da AMD nei loro laboratori tedeschi.

Spesso si assume che il valore misurato per una macchina con una certo processore sia valido per tutte le macchine con quel processore, mentre in realtà HS06 misura le prestazioni di una macchina in una certa configurazione complessiva del hardware. La misura è influenzata da diversi fattori tra cui i più significativi sono la versione del compilatore, del kernel e la configurazione di memoria (quanta memoria per core, quanti banchi di memoria, clock e bandwidth dei banchi di memoria). Le misure possono anche essere influenzate dal chipset della motherboard e dalla temperatura ambientale.

L'influenza della memoria è molto importante perché può influire pesantemente sul rapporto Euro/HS06. Le prestazioni in termini di HS06 diminuiscono al diminuire della memoria perché aumentano le probabilità di swapping. Per questo motivo nelle nostre misure lo swapping viene disabilitato. In questo modo il benchmark o gira senza problemi o fallisce.

Per dare un esempio su di una macchina con processori 5520 la misura di HS06 varia in questo modo con il variare dei moduli di memoria installati.

1) 32GB (8*4GB) → 109.11 HS06
2) 40GB (10*4GB) → 113.48 HS06
3) 48GB (12*4GB) → 115.70 HS06

4.1 La curva del 5650

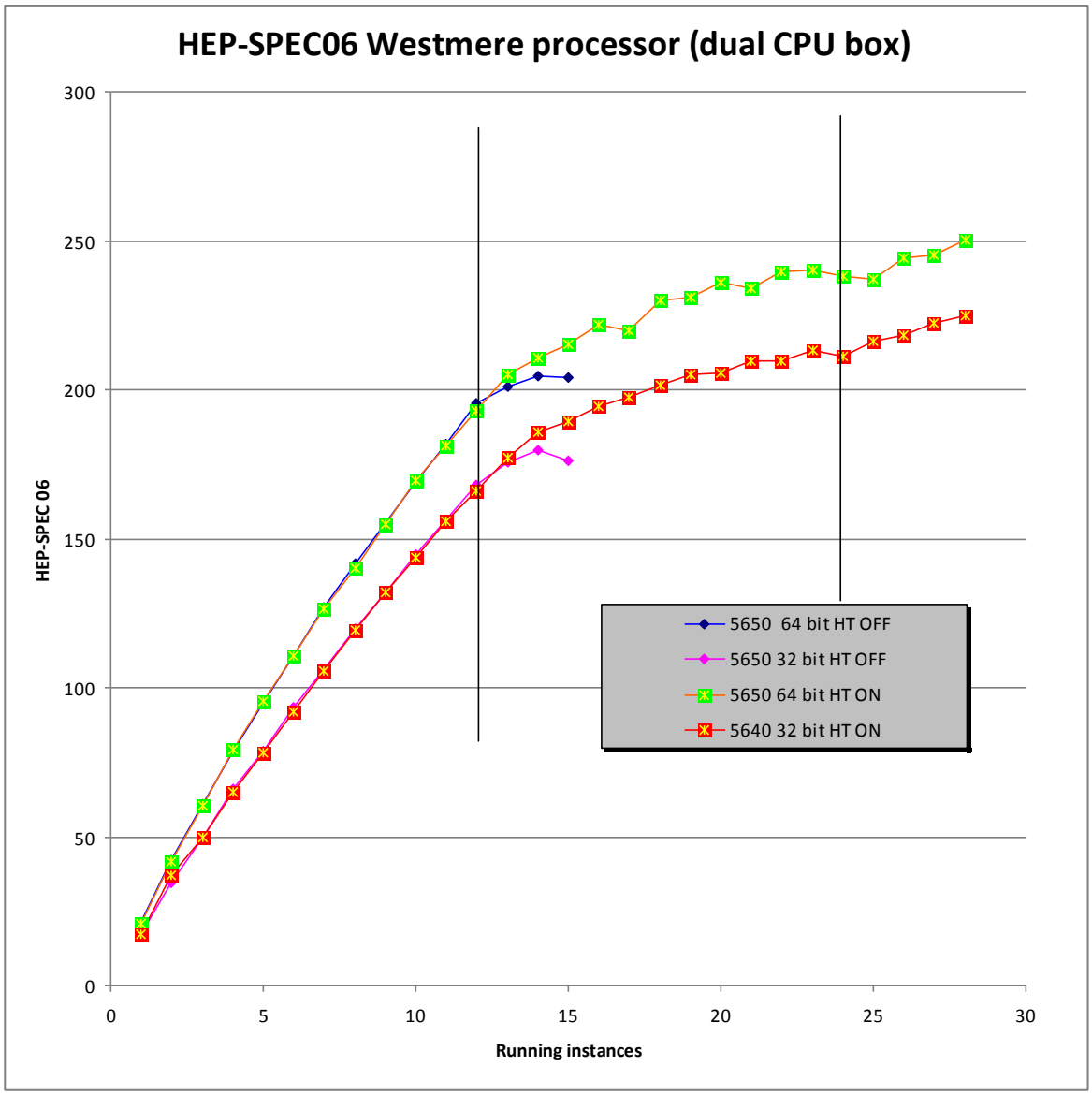
La macchina con il doppio 5650 è stata misurata avere un throughput di 166.97 +/- 0.48 HS06 con HT-ON e 167.15 +/- 0.62 con HT-OFF. Se volessimo considerare la macchina come se avesse 24 CPU con HT-ON raggiungiamo il valore di 211.56 +/- 2.08.

La macchina aveva 48 GB di memoria per cui non ci sono stati problemi nel girare con 24 istanze contemporanee. Ovviamente nell'uso di produzione l'esperimento deve decidere quale è il valore corretto di Gigabyte per core.

Abbiamo misurato anche HS06 a 64 per avere un'idea del miglioramento del codice a 64 bit per gli esperimenti che hanno già fatto la transizione.
In questo caso con HT-ON abbiamo 192.43 +/- 0.76 con 12 istanze e 238.83 +/- 1.12 con 24 istanze parallele.

Nel grafico si possono vedere le quattro curve. In alto quelle a 64 bit e in basso quelle a 32 bit. Quelle che arrivano oltre il segmento verticale delle 24 istanze parallele (24 cpu logiche) sono con l'HT-ON mentre quelle che si fermano poco dopo il segmento verticale dei

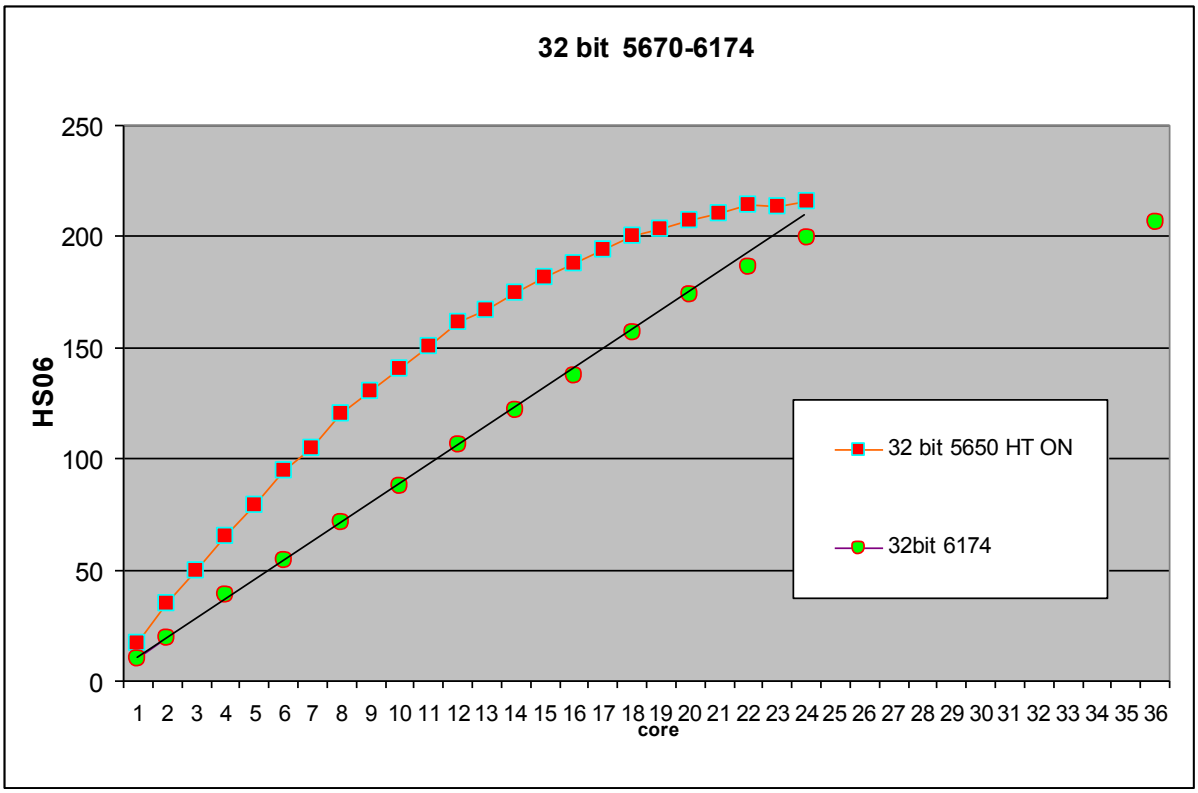
12 core fisici sono con HT-OFF.



4.2 La curva dell'Opteron 6174

Non siamo riusciti a ricostruire la stessa curva in dettaglio per il processore AMD Magny-Cours ma riportiamo qui sotto una curva parziale confrontata con un processore Intel 56xx. Si vede che l'andamento è molto più lineare perché le cpu logiche AMD corrispondono a veri e proprio core fisici mentre i processori con HT hanno una incremento quasi lineare fino al numero di core fisici e un andamento lineare con pendenza inferiore quando si usano le risorse della seconda cpu logica all'interno dei core fisici.

Il processore AMD Opteron 6174 a 2.2 GHz ha fornito 199.68 HS06 ed oltre 233.87 con la versione HS06 a 64 bit.



5 RIFERIMENTI

[1] A comparison of HEP code with SPEC benchmarks on multi-core worker nodes. Michele Michelotto et al 2010 J. Phys.: Conf. Ser. 219 052009

[2] SPEC http://www.spec.org