ISTITUTO NAZIONALE DI FISICA NUCLEARE
Sezione di Padova

INFN/CCR-08/02
22 Dicembre, 2008

CCR-29/2008/P
Ottobre 2008
Versione 1.0.1

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 i nodi di calcolo (*Worker Node*) nei prossimi mesi.

La valutazione delle prestazioni dei processori in questo momento di transizione tra SPEC INT 2000 e un altro *benchmark,* probabilmente della suite SPEC CPU 2006, risulta alquanto problematica poiché vengono usati vecchi e nuovi *benchmark*, con dati in parte pubblicati e in parte misurati.

Il documento si conclude con una tabella in cui si evidenziano le prestazioni delle macchine attuali e di quelle delle generazioni precedenti in termini di SI2k e SPEC INT 2006 in attesa della scelta di WLCG come *benchmark* di riferimento per il futuro.

Infine sono riportati i prezzi di alcuni medi acquisti di *Worker Node* (taglia tipica dei Tier2) degli ultimi mesi nell'intento di valutare i prezzi per gli acquisti futuri (2009).

Pubblicato da ***SIS–Pubblicazioni***
Laboratori Nazionali di Frascati

1 INTRODUZIONE

Lo scopo di questo documento è presentare una panoramica dei processori disponibili sul mercato per i nodi di calcolo (*Worker Node*) nei prossimi sei e dodici mesi. Si cercherà inoltre di prevedere i prezzi per acquisti tipici da Tier2 (10-30 box).

Saranno presentate le attuali tecnologie dei semiconduttori, l'influenza della legge di Moore sul *clock,* la dimensione delle *cache* e i consumi energetici. Il tipico *worker node* sarà ancora *dual processor* con processori *dual quad core* e 2 GB di memoria per ogni *core*.

Le configurazioni a *blade,* avendo ormai gli stessi prezzi delle configurazioni "pizza box", risultano essere più convenienti in termini di costi totali di esercizio grazie alla riduzione dei consumi del 20 – 30%. Quanto affermato risulta però vero solo acquistando un numero di *blade* tale da ammortizzare il costo del *crate*. Sono molto diffuse anche le configurazioni *twin*: due macchine gemelle che condividono box esterno e alimentatore.

2 UNITÀ DI MISURA

L'unità di misura delle prestazioni è attualmente un aspetto molto critico. Le richieste degli esperimenti e le potenze promesse dai centri di calcolo Tier1/Tier2 sono in termini di SI2K o kSI2K, abbreviazioni comunemente usate per i *benchmark* "Integer" della suite CPU 2000 pubblicata da SPEC.

Da un paio di anni CPU 2000 è stata sostituita da SPEC con la suite CPU 2006 per obsolescenza tecnologica. Gli avanzamenti tecnologici delle CPU, delle loro architetture di *cache* e delle memorie RAM dei sistemi hanno reso inaffidabili le valutazioni di performance in termini di CPU 2000.

Ad 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 di tipo *CPU bound,* mentre utilizzando gli SI2K pubblicati dal sito *www.spec.org* il primo tipo di processore ha un *rating* doppio rispetto al secondo.

Questa enorme differenza è dovuta soprattutto all'uso nei programmi HEP del compilatore **gcc** invece di compilatori commerciali quali Intel o Pathscale. Confrontando ad esempio SI2K calcolato sui vecchi processori Xeon e sui nuovi Woodcrest e Clovertown osserviamo una perdita di prestazioni sino al 55-65% del gcc (compilando con ottimizzazione non molto spinta, quella solitamente usata nei codici LHC) rispetto a **icc**.

TAB. 1: Il compilatore gcc fornisce valori di SPEC INT sensibilmente inferiori 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%

Per ovviare a questo inconveniente il CERN da qualche tempo richiede che i partecipanti ai *tender* per l'acquisto dei *Worker Node* forniscano gli SPEC INTmisurati con il

compilatore **gcc** e con il seguente insieme di ottimizzazioni: ***–O2 –fPIC –pthread***.

In questo modo diventa possibile calcolare SI2K per processori di nuova generazione per i quali SPEC non accetta più la pubblicazione dal 2006 rendendo di conseguenza impossibile qualsiasi confronto tra SPEC misurato e SPEC pubblicato.

Chiameremo d'ora in poi gli SPEC INT calcolati con il *tuning* proposto dal CERN con il nome SPECINT-CERN o SI2K-CERN.

Abbiamo visto come i vaolori delle prestazioni calcolati con il compilatore gcc fossero ormai la metà di quelli pubblicati per le ultime macchine su cui si potessero fare confronti. In questo modo si rischiava di compiere errori del 100% nell'acquisto di macchine basandosi, come spesso si era fatto in passato, su SPECINT misurati invece che su SPECINT pubblicati.

La soluzione "pro tempore" proposta dal Management Board di WLCG è stata calcolare un valore di SPECINT equivalente a quello del 2001.

Questa soluzione nasce dalla proposta del Tier1 tedesco di Karlsruhe: misurando i valori di SPECINT 2000 per macchine del 2001 si ottenevano valori pari a circa l'80% di quelli pubblicati, usando però un *tuning* di gcc più spinto rispetto a quello CERN. Il *tuning* di GridKa è infatti dato da "**gcc –O3 –unroll-loops –march=$ARCH"** dove "ARCH" è l'architettura della macchina.

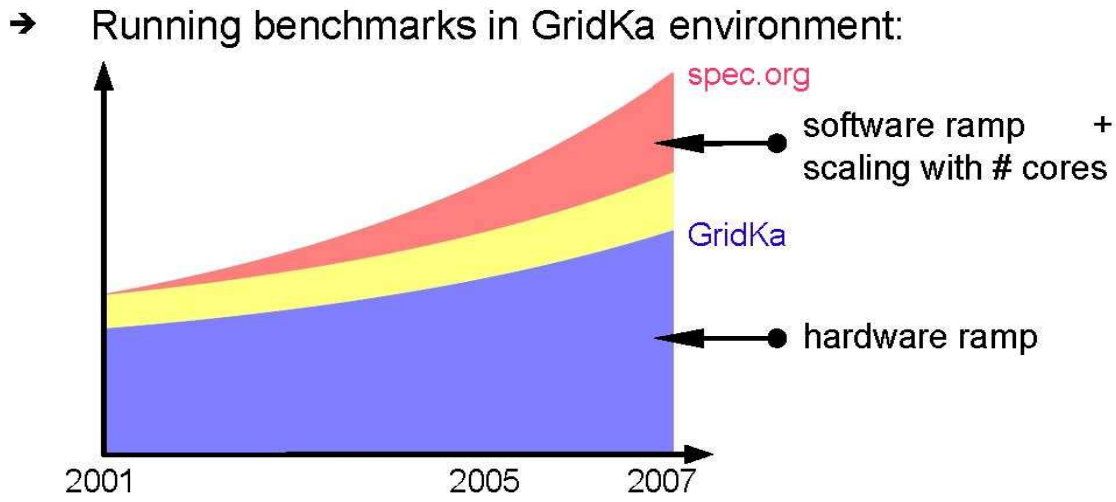

FIG. 1: Scostamento tra i valori SPEC pubblicati e i valori misurati da GridKa

GridKA quindi proponeva di usare come valore di SPECINT per una macchina quello misurato con il loro *tuning*, aumentato del 25%, rinormalizzando in questo modo i valori misurati ai valori del 2001.

Ripetendo lo stesso ragionamento ma con il *tuning* CERN (***gcc –O2 –fPIC –pthread***) si ottengono valori di SI2K-CERN misurati inferiori (circa il 60-70% di quelli pubblicati) per cui la proposta del Management Board di LCG è stata di usare come valori di riferimento di una macchina quelli misurati con il *tuning* CERN aumentati del 50%.

Chiameremo questi SI2K-CERN rinormalizzati con il nome di SPECINT 2000 LCG o SI2K-LCG.

Tuttavia ritengo necessario passare al più presto all'uso di Spec CPU int 2006 perché è meno sensibile di SI2K alla dimensione delle *cache* di secondo livello ed è pensato per occupare circa 1 GB di RAM per ogni *core* contro i 200 MB/*core* di SI2K.

3 IL TIPICO 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 ha ormai quattro *core*, permettendo quindi di processare otto *job* in parallelo.

L'alternativa è la configurazione a *blade* in cui i box vengono installati in verticale in cestelli (*crate*) dotati di alimentatori ridondati ad alta efficienza. Il vantaggio della configurazione a *blade* sta nel fatto che tutti i nodi continuano a funzionare anche se uno degli alimentatori si rompe.

Invece nel caso in cui si rompa l'unico alimentatore di un nodo con 8 *core* causano la perdita di 8 code batch (infatti macchine di questo tipo, usate con funzioni di server andrebbero dotate di due alimentatori per ridondanza).

Inoltre la configurazione a *blade* può avere un unico accesso fisico locale (permettendo di risparmiare i costosi *switch* KVM), un unico accesso IPMI e spesso è dotata di *switch* Ethernet integrati.

Lo svantaggio è rappresentato soprattutto dal fatto che si devono acquistare un numero minimo di *blade* per ammortizzare l'acquisto dell'infrastruttura (alimentatori, *crate*, *switch*).

Dal punto di vista dei costi le configurazioni a *blade*, a cestelli pieni, risultano avere un costo maggiore del 10 – 20% rispetto alle configurazioni 1U. Tuttavia il costo totale di esercizio dei *blade* viene pubblicizzato (IBM, HP, Supermicro) come più conveniente poiché gli alimentatori delle macchine 1U hanno efficienza energetica quasi pari al 75% mentre per quelli dei *blade* è superiore al 90%.

Esempio: Con un carico di 500 Watt un box 1U il cui alimentatore ha un'efficienza del 75% avrebbe bisogno di 667 Watt in ingresso mentre un *blade* richiederebbe solo 538 Watt, portando ad un risparmio di 129 Watt: in un anno e a pieno carico corrispondono a 1130 kWh. Prendendo come costo del kWh 0.15 Euro si ha un risparmio di 169.50 Euro/anno.

Sui tre anni di vita di una macchina, acquistando un cestello di 10 nodi, si avrebbe un risparmio di 5085 Euro, senza contare il risparmio da imputarsi al minor consumo dovuto a 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 sulle minori richieste di impianti tecnologici infrastrutturali (UPS, gruppi elettrogeni, condizionatori) e di conseguenza sul minor consumo elettrico degli impianti stessi.

4 TWIN SERVER

Un'altra possibilità è data dalle configurazioni *twin*. In un *case* 1U vengono affiancate due *motherboard* simili a quelle dei *blade*, molto strette e lunghe: l'equivalente di due *blade* posti in orizzontale. Il vantaggio rispetto ai *blade* è il risparmio dei costi dell'infrastruttura del cestello. Si hanno all'atto pratico due server nello stesso *case* con un unico alimentatore.

Chi scegliesse configurazioni di questo tipo deve prestare molta attenzione alla

questione della dissipazione termica. Infatti queste macchine hanno alimentatori da 900 Watt con consumi prossimi ai 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 per machine a due vie disponibili in questo momento sono i seguenti:

Clovertown: la serie *quad core* **Intel 53xx**, *clock* da 1600 a 3000 MHz. Stessa *cache* del Woodcrest 51xx: come avere due *chip* Woodcrest affiancati nello stesso *chip enclosure* e quindi in un *socket*. La *cache* di secondo livello è di 4 MB per ognuno dei due "die" e i consumi vanno dai 50 W ai 120 W.

Harpertown: la serie *quad core* **Intel 54xx**, *clock* da 2000 a 3200 MHz. Molto simili ai Clovertown, sono costruiti in tecnologia a 45 nm. Hanno *cache* da 6 MB invece che da 4 MB. Consumi dai 50 W agli 80 W. Solamente il 5482 a 3200 MHz consuma 150 W.

Barcelona: la serie *quad core* **AMD 23xx**. Arrivato sul mercato a volumi con un ritdardo di almeno sei mesi rispetto alle aspettative. Si tratta di un vero processore *quad core* nel senso che sul singolo *die* sono stampati quattro *core*. Costruiti in tecnologia a 65 nm hanno *clock* da 1900 a 2600 MHz. Sono caratterizzati da una piccola *cache* L2 di 512 KB completamente dedicata al *core* e da una *cache* di terzo livello da 2 MB condivisa dai quattro *core* presenti nel processore.

I processori *dual core* **Intel Xeon Woodcrest** e **AMD Opteron** non sono più competitivi.

6 PROCESSORI PER I PROSSIMI ACQUISTI

Per l'inizio del 2009 dovrebbero essere disponibili i nuovi processori Intel "**Nehalem**" che assomigliano molto agli AMD Barcelona: tre livelli di *cache*, *memory controller* integrato, quattro *core* nello stesso *die*, *cache* di maggiore dimensione. Le prestazioni sono sconosciute ma si parla di un aumento del 40% - 70% a parità di *clock*. Non è chiaro se questo aumento avverrebbe per ogni *job* o considerando la possibilità di eseguire due *thread* in parallelo sullo stesso *core*. Di solito il *multi-threading* non ha mai dato vantaggi evidenti nel codice HEP ma essendo questa una nuova esecuzione dell'*hyperthreading* va studiata.

Se però risultasse possibile eseguire due *job* HEP sullo stesso *core* è probabile che le esigenze in termini di memoria raddoppino. La memoria dovrebbe essere DDR3 ma non più *fully buffered*.

Dovrebbe essere disponibile tra Q4 2008 e Q1 2009, quindi per i prossimi acquisti dei Tier2, con il nome in codice di "**Gainestown**": per le versioni *dual processor* con 256 KB di *chache* L2 e 6 – 8 MB di *cache* L3 condivisa, *clock* da 1.86 GHz per il 5502 e 2.97 GHz per l'X5570, consumi compresi tra 60 e 95 Watt per processore.

La risposta AMD è il processore "**Shangai**" disponibile da Novembre 2008 con 2 MB di *cache* L2, 6 MB di *cache* L3, tecnologia a 45 nm, *clock* da 2.1 GHz per l'Opteron 2372 HE e fino a 2.8 GHz per il 2386 SE. I consumi sono compresi tra i 55 Watt (HE) e i 105 Watt (SE). Altri modelli usciranno nel corso dell'anno fino a febbraio del 2009.

7 FINE 2009 E OLTRE

Le novità nel 2009 saranno l'aumento della dimensione delle *cache* di secondo e terzo livello e il passaggio alla tecnologia a 32 nm: quest'ultima in particolare permetterà un maggior numero di *transistor* portando quindi un aumento ulteriore del numero dei *core.*

Il processore a 6 *core* di AMD dovrebbe chiamarsi **"Istanbul"** e funzionare sulle stesse schede madri "Socket 1207" utilizzate dai processori "Shangai".

Nel H1 2010 invece AMD presenterà il nuovo *socket* G34 per i processori della famiglia "Sao Paulo" con 6 *core* e *memory controller* DD3 e a seguire il processore "Magny Cours" con 12 *core* ottenuti con due *die* "Sao Paulo" affiancati nello stesso *package*.

Il successore del Gainestow sarà il **Westmese**, un processore DP a 32 nm con 4 o 6 *core* (quindi 8 o 12 *thread*), *cache* L3 raddoppiata a 12 MB e raddoppio dei canali verso la memoria DD3 (Q4 2009, H1 2010).

Il primo processore a sei *core* **Intel** è già disponibile ma solo nella versione MP (per macchine con almeno quattro processori) e non nelle tradizionali configurazioni a due vie usate solitamente in ambiente HEP. Si tratta dell'ultima evoluzione del *core* di tipo Penryn ed è l'equivalente della serie 54xx nella famiglia MP, si chiama infatti **74xx**.

8 PRESTAZIONI DEI WORKER NODE ATTUALI

Nella tabella seguente sono riportate le prestazioni misurate per i processori attuali e per quelli della scorsa generazione usando SI2K-CERN e SPECINT 2006 (sempre gcc e *tuning* CERN).

Ribadisco quando detto sopra. Se dovete comprare SI2K per rispondere a richieste LCG (quindi esperimenti LHC) questi SI2K vanno moltiplicati per 1.5 al fine di avere il *rating* ufficiale LCG della macchina.

L'ultima novità del gruppo di *benchmarking* HEPIX è che alcuni membri del gruppo spingono per usare un nuovo *benchmark* ottenuto prendendo solo i *benchmark* scritti in C++ tra tutti quelli di CPU 2006 che concorrono ai risultati SPECINT e SPECFP. Questo nuovo *benchmark* ha il vantaggio di avere più codice *floating point* rispetto a INT 2006 che praticamente non ne ha. La situazione in questo momento è ancora fluida per cui non userò questa nuova unità di misura nel resto del documento.

I numeri riportati in seguito si riferiscono alle prestazioni dell'intera macchina sotto carico. Invece di usare SPECINT RATE (2000 o 2006), come proposto da SPEC, si eseguono tante copie di SPECINT e infine si somma il risultato ottenuto su tutti i *core*.

Per CPU 2006 il numero finale è leggermente più elevato rispetto a quello ottenuto con SPECINT RATE: nel primo caso non appena un *core* ha finito viene lanciato il *benchmark* successivo mantendo tutti i *core* sempre in funzione (come su di una *batch farm*) mentre con SPECINT RATE è necessario attendere ogni volta che tutti i *core* abbiano finito un "sottobenchmark" e inoltre viene considerato come tempo di esecuzione il tempo in cui finisce il più lento dei diversi *job* in esecuzione parallela.

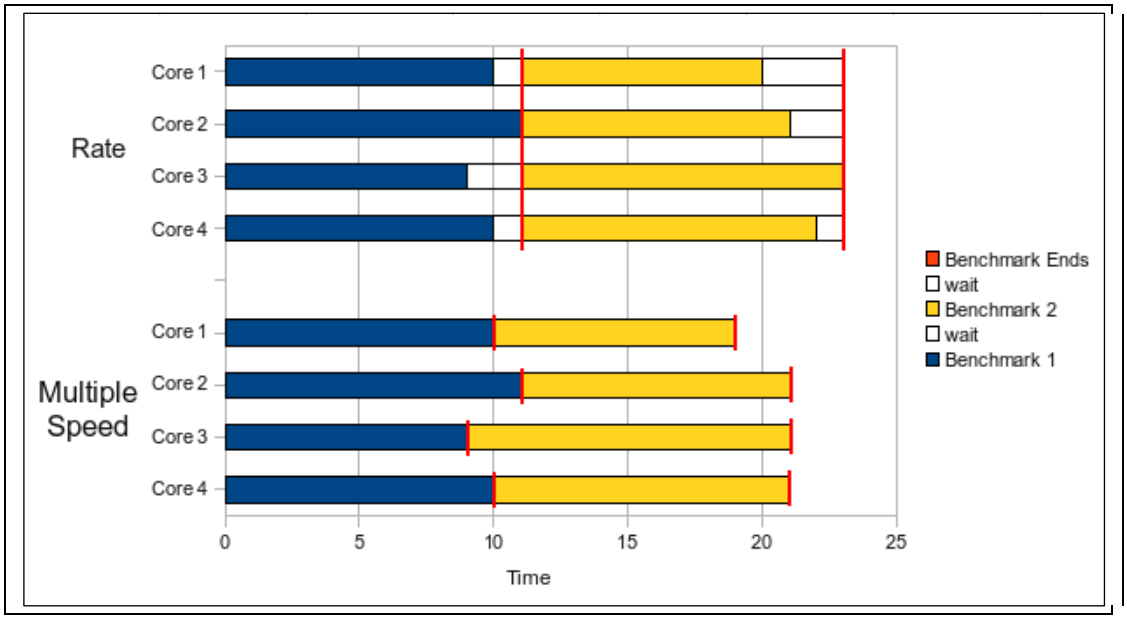

FIG. 2: Differenza tra uso di SPEC INT Parallelo e SPEC INT RATE.

Come si può notare dalla figura più *core* ci sono in una macchina e più aumenta la varianza delle prestazioni tra un *core* e l'altro. Quindi una macchina che misura in modalità RATE sarà sempre più lenta di una che misura in modalità MULTIPLE SPEED.

Nella tabella di seguito si riportano i valori di SPECINT 2000 CERN e SPECINT 2006 CERN misurati con gcc su macchine con due processori (non rivalutati del 50%). La memoria è sempre di 2 GB per *core*: quindi 16 GB per macchine 2x4 *core* e 8 GB per macchine 2x2 *core*.

NB. Ripeto quanto detto prima: SPECINT 2000 CERN significa: SPEC CPU 2000 misurato con "gcc –O2 –fPIC –pthread" sommato sul numero dei *core* presenti nella macchina NON rivalutato del 50%.

SPECINT 2006 CERN come sopra ma con SPEC CPU 2006.

Per indicare lo SPECINT 2000 CERN rivalutato del 50% secondo le indicazioni LCG userò la sigla SI2K-LCG.

Sono da notare i tre diversi valori ottenuti per il 5410 che riflettono lo *spread* risultante dalla misurazione di un certo *benchmark* su macchine che pur avendo lo stesso *chip* presentano caratteristiche differenti (tipo di memoria, *motherboard*, *bios*).

TAB. 2: Valori di SI2K CERN e SI2006 CERN per processori in uso attualmente

Processore	Clock - Memoria	SPECINT 2000 CERN	SPECINT 2006 CERN
Intel Nocona	2.8 GHz – 2GB	1501	11.06
Amd Opteron 275	2.2 GHz – 4GB	4133	28.76
Intel Woodcrest 5150	2.66 GHz – 8GB	5675	36.77
Intel Woodcrest 5160	3.0 GHz – 8GB	6181	39.39
Amd Opteron 2218	2.6 GHz – 8GB	4569	31.40
Intel Clovertown 5345	2.33 GHz 16GB	9462	60.89
Intel Harpertown 5410	2.33 GHz 16GB	10556	64.78
Intel Harpertown 5410	2.33 GHz 16GB	11850	73.32
Intel Harpertown 5410	2.33 GHz 16GB	11164	68.20
Amd Barcelona 2352 b2	2.10 GHz 16GB	8488	56.23
Amd Barcelona 2360 b3	2.50 GHz 16 GB	9939	68.50
Intel Harpertown 5430	2.66 GHz 16GB	12259	73.24

9 I PREZZI

Il prezzo di una macchina *quad core* con 16 GB di memoria può variare molto. Con lo stesso tipo di processore ho visto prezzi compresi tra i 1800 e i 2500 Euro (IVA esclusa). Per processori con *clock* più elevato si oltrepassano i 3000 Euro.

Le macchine in configurazione *twin* sono spesso più convenienti. Prima di tutto perché due *worker node* condividono *case*, slitte, dispositivi di I/O e alimentatore.

Da una ricognizione dei prezzi praticati all'INFN nel 2008 sono risultate altrettanto convenienti le configurazioni a *blade* ma solo nel caso in cui si riescano a riempire completamente i cestelli.

Vediamo alcuni confronti di prezzo / prestazioni per determinati acquisti dell'INFN a fine 2007 (i primi quattro) e nel 2008 (in grassetto nella tabella). Solo per configurazioni *quad core* e acquisti totali di almeno una ventina di *core*.

I prezzi sono IVA inclusa. Il rapporto prezzo / SPEC INT è nelle ultime due colonne. Nella penultima c'è il prezzo in Euro per migliaia di SPEC INT 2000 LCG. Questo è stato calcolato dividendo il valore in Euro per la valutazione in kSI2k misurati con il *tuning* CERN e rivalutati del 50% secondo le direttive dell'LCG MB. L'ultima colonna invece è relativa agli Euro per SPEC INT 2006 sempre misurati con il *tuning* CERN.

TAB. 3: Rapporto prezzo / prestazioni per acquisti recenti nell'INFN

tipologia	Numero di WN	Cores totali	Processore	Euro per kSI2k LCG	Euro per SI2006
Blade	29	232	2 x 5430	192	48
Blade	10	80	2 x 5430	232	58
Blade	14	112	2 x 5420	133	33
Blade	10	80	2 x 5450	225	56
Blade	**16**	**128**	**2 x 5430**	**147**	**37**
Blade	**16**	**128**	**2 x 5430**	**135**	**34**
Twin	**14**	**112**	**2 x 5410**	**164**	**41**
Twin	**6**	**48**	**2 x 5410**	**140**	**35**
Twin	**6**	**48**	**2 x 5410**	**157**	**39**
Twin	**22**	**176**	**2 x 5440**	**137**	**34**
1 U	**5**	**40**	**2 x 5410**	**194**	**49**

Vediamo come i prezzi nel **2008** (in grassetto) siano compresi tra i **137** e i **157 Euro/kSI2k-LCG** e tra i **34** e i **39 Euro/SI2006**. Prezzi migliori sono stati ottenuti solo nelle gare del Tier1.

Per il futuro è lecito aspettarsi un'ulteriore riduzione almeno del 20% per i prossimi tre mesi e del 30% nei mesi successivi per i seguenti motivi:

- discesa del prezzo delle memorie
- nuovi processori leggermente più costosi ma più performanti
- netta discesa del prezzo dei processori attuali (per esempio la serie 54xx)

10 LA CURVA DEI PREZZI

Ho chiesto ad un fornitore il prezzo attuale per alcune macchine in formato 1U, *twin* e *blade*. Risultano più convenienti le macchine in formato *twin*, ma la differenza è minima. In questo momento sembrano più convenienti le macchine con processore Intel.

È da notare che per Intel risultano più convenienti le macchine con un *clock* intermedio mentre per AMD le macchine con un *clock* più basso.

Per confronto sono state considerate anche le macchine relative agli acquisti dei Tier1 e dei Tier2 analizzati nelle tabelle precedenti. Come si può notare diversi acquisti sono stati effettuati a prezzi migliori rispetto ai più bassi che ho ricevuto. Molto probabilmente quindi non sono esattamente i prezzi migliori per acquisti di taglia Tier2.

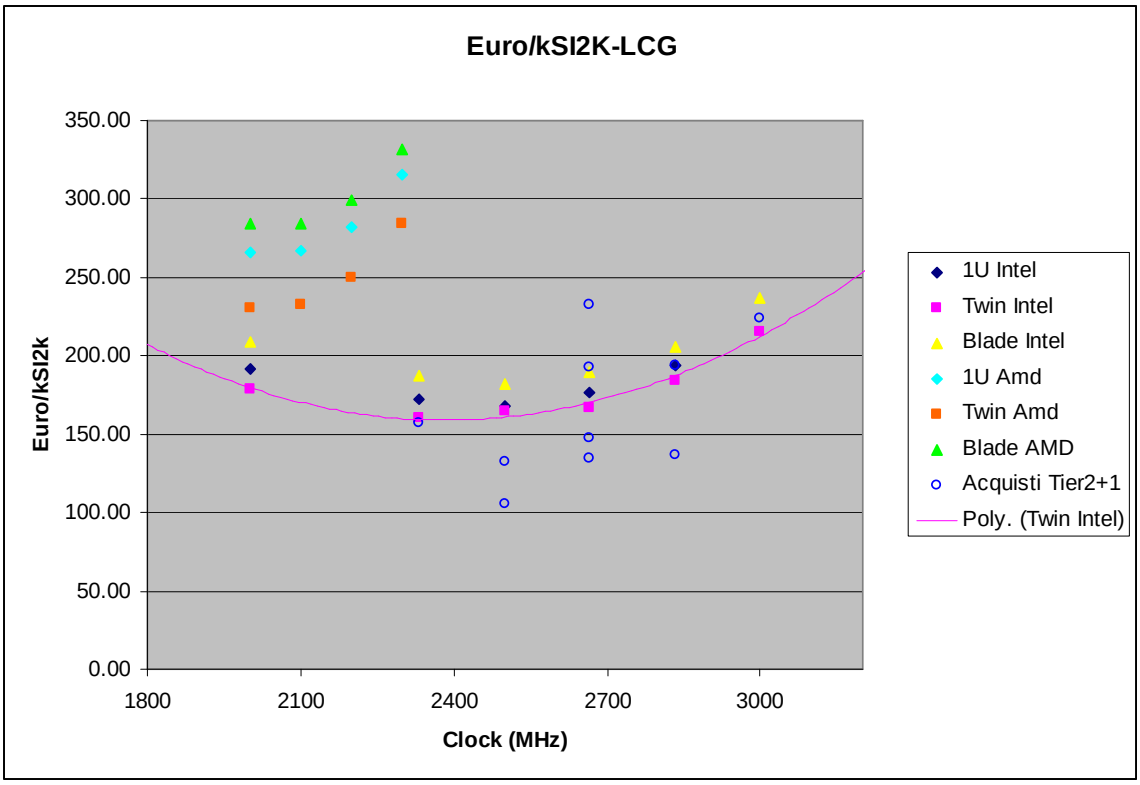

FIG. 3: Curva prezzo / prestazioni per diverse tipologie di WN in rapporto al *clock*.

11 RINGRAZIAMENTI

Vorrei ringraziare Mauro Morandin, Massimo Masera, Roberto Stroili, Luciano Barone, e Giampaolo Carlino che mi hanno fornito dati indispensabili per la creazione delle tabelle degli acquisti 2007 - 2008.

12 RIFERIMENTI

HEPiX CPU Technology tracking
http://hepix.caspur.it/afs/hepix.org/project/ptrack/

SPEC http://www.spec.org/