DA 2 MILA A 2 MILIARDI

DI GIANCARLO MAGNAGHI

Feedback: redazione-cbr@edizionibig.it

È la crescita del numero di transistor in un circuito integrato, nell'evoluzione della tecnologia e delle architetture dei microprocessori in quasi 40 anni

Da quando, nel 1971, Federico Faggin creò Intel 4004, il primo microprocessore a 4 bit basato su un circuito integrato con circa 2000 transistor, la tecnologia dei microprocessori si è evoluta seguendo la legge di Moore, enunciata nel 1965 da Gordon Moore, fondatore di Intel: essa prediceva che ogni 18-24 mesi sarebbe raddoppiato il numero di transistor incorporati in un circuito integrato. Infatti, come si vede dalla tabella 1, dopo quasi quarant'anni di continui sviluppi il microprocessore Itanium quad core di Intel ha superato la barriera dei 2 miliardi di transistor su un chip: un milione di volte la capienza del 4004. Questo mare di transistor è stato utilizzato in vario modo: innanzitutto per aumentare il parallelismo della CPU e delle memorie , che è cresciuto da 4 a 8, 16, 32, 64 e 128 bit; poi per aumentare la potenza del set di istruzioni macchina, le logiche di controllo, i dispositivi di accelerazione dell'esecuzione e le funzioni di sicurezza.

Le evoluzioni principali nell'architettura dei microprocessori sono derivate dalla necessità di aumentare le prestazioni (vedi tabella).

Microprocessore	Bit	Anno	Transistor	Produttore
4004	4	1971	2,25 K	Intel
8080	8	1974	5 K	Intel
Z80	8/16	1976	6 K	Zilog
8088	8/16	1979	29 K	Intel
80286	16	1982	134 K	Intel, Amd
80386	32	1985	275 K	Intel, Amd
80486	32/64	1989	1,2 M	Intel, Amd
Pentium	32/64	1993	3,1 M	Intel
Pentium III	32/64	1999	9,5 M	Intel
Athlon	32/64	1999	22 M	Amd
Pentium IV	32/64/128	2000	40 M	Intel
Opteron	32/64/128	2003	100 M	Amd
Itanium 2	32/64/128	2004	220M	Intel
Opteron quad core	32/64/128	2006	460 M	Amd
Core 2 Quad	32/64/128	2006	582 M	Intel
Xeon MP 6 core	32/64/128	2008	1.900 M	Intel
Itanium quad core	32/64/128	2008	2.000 M	Intel

Tabella 1 - Evoluzione dei microprocessori con architettura Intel

Misura e incremento delle prestazioni

Esistono diversi modi per misurare le prestazioni di un sistema di elaborazione. La misura più utile agli effetti pratici è "il tempo di esecuzione necessario a completare un determinato programma", che dipende da vari fattori, tra cui le prestazioni del microprocessore, del sottosistema di memoria, del sottosistema di I/O, nonché dal tipo di programma eseguito.

Se limitiamo l'analisi alle prestazioni del microprocessore, valgono le relazioni:

$$T_{exec} = N_i / IPS \quad e \quad IPS = F_{clock} \times IPC = F_{clock} / CPI$$

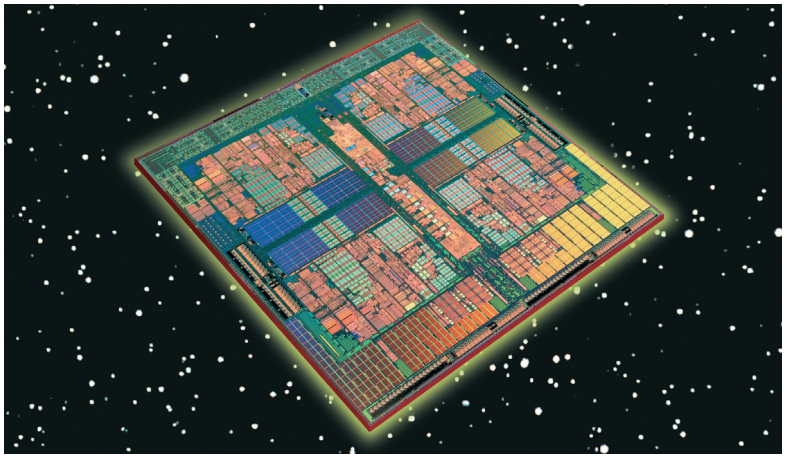
dove:

Texec	=	Tempo di esecuzione del programma
IPS	=	Numero di istruzioni per secondo
Ni	=	Numero di istruzioni del programma
IPC	=	Numero di istruzioni per ciclo di clock
Fclock	=	Frequenza di clock del processore
CPI	=	cicli di clock per istruzione

Da queste formule risulta chiaro che è possibile aumentare la velocità di esecuzione agendo su tre fattori: aumentare la frequenza di clock (Fclock); ridurre il numero di istruzioni macchina (Ni), a parità di lavoro svolto; aumentare il numero di istruzioni per ciclo di clock (IPC) o, in altre parole, ridurre il numero di cicli di clock per istruzione (CPI).

Aumento della frequenza di clock. Inizialmente, si è agito soprattutto per aumentare Fclock, facendo leva sulla miniaturizzazione,

che fa rimpicciolire i processori e migliora la velocità, i consumi e la dissipazione termica, passando dai 108 KHz dell'Intel 4004 ai quasi 4 GHz dei microprocessori odierni. Ma la tecnologia VLSI disponibile non permette di aumentare all'infinito la densità (le attuali geometrie di 45 nanometri sono vicine ai limiti della tecnologia del silicio ed è sempre più difficile dissipare il calore prodotto), inoltre, al crescere della frequenza di clock emergono altri problemi legati alle prestazioni del sottosistema di memoria, poiché la velocità di accesso alla memoria è cresciuta meno della velocità dei processori. Per aggirare questo problema si usano piccole memorie di transito veloci (memorie *cache*), organizzate secondo livelli gerarchici (Cache L1, L2, L3), che sfruttano il *principio di località* per ridurre il tempo di accesso medio.



AMD Phenom Quad-Core processor

Riduzione del numero di istruzioni macchina. Mentre i primi microprocessori disponevano di poche decine di istruzioni, i moderni microprocessori sono dotati di centinaia di istruzioni, un grande numero di registri e di unità aritmetiche e un elevato livello di parallelismo (fino a 128 bit). Fino agli anni '80 ci fu la corsa all'aumento del numero di istruzioni macchina, che portò alle cosiddette architetture CISC (Complex Instruction Set Computer), caratterizzate da una eccessiva complicazione delle unità di controllo micro programmate, a cui l'industria reagì introducendo le macchine RISC (Reduced Instruction Set Computer) che privilegiano la velocità di esecuzione rispetto alla potenza delle istruzioni macchina.

Riduzione dei cicli di clock per istruzione (CPI). Nei primi microprocessori, per eseguire un'istruzione macchina erano necessari anche 20 cicli di clock per i trasferimenti tra registri e da/verso la memoria di lavoro. Per risolvere questo problema si introdussero vari accorgimenti che permettono l'esecuzione in parallelo di più istruzioni, eseguendo negli anni '90 il *downsizing* a livello chip degli accorgimenti adottati negli anni '70 sui supercalcolatori e sfruttati a partire dagli anni '80 su minicomputer e PC (coprocessori matematici e grafici, pipeline, architetture multibus). Per esempio, l'architettura Harvard permette un parallelismo di esecuzione superiore a quello dell'architettura classica di Von Neumann (in cui la CPU legge istruzioni e dati dalla memoria utilizzando lo stesso bus), poiché separa i dati dalle istruzioni e parallelizza le operazioni di lettura e scrittura della memoria. Praticamente tutti i moderni processori dividono la memoria "cache dati" dalla "cache istruzioni" per accedere in parallelo alle due cache e migliorare le prestazioni. Ma l'accelerazione più rilevante si ottiene dotando il microprocessore di una "Pipeline", che consente di eseguire in parallelo le fasi elementari di varie istruzioni consecutive. Una pipeline ideale consente di eseguire un'istruzione per ogni ciclo macchina (CPI=1). Questo però in pratica non si verifica mai, perché quando si incontra un'istruzione di salto la pipeline deve essere reinizializzata e si perdono cicli macchina. Per alleviare questo problema, sono stati introdotti dispositivi di predizione dei salti e di esecuzione speculativa. Si possono ottenere risultati migliori con la tecnica dell'*esecuzione super scalare*, che permette di lanciare più istruzioni contemporaneamente (CPI<1), ma richiede che all'interno della CPU molte risorse siano duplicate. In pratica si hanno più pipeline parallele, ciascuna delle quali elabora una diversa operazione, mentre una logica di controllo provvede a smistare le istruzioni alle pipeline e a raccogliere i risultati.

Un processore con una singola pipeline (che decodifica solo un'istruzione per volta) viene detto *processore scalare* mentre i processori con più pipeline (che possono eseguire più di un'istruzione in un ciclo di clock)vengono detti *superscalari*. Un microprocessore con architettura superscalare supporta il calcolo parallelo su un singolo chip, permettendo prestazioni molto superiori a parità di clock rispetto ad una CPU ordinaria. Questa caratteristica è posseduta più o meno da tutti i microprocessori *general purpose* prodotti dalla fine degli anni '90. La tecnologia *Hyper-Threading* (HT Technology), introdotta da Intel nel 2002, permette ai microprocessori di eseguire task in parallelo gestendo più "thread" in un singolo processore. Il processore "appare" al software come se fosse composto da due processori

fisici e utilizza in modo più efficiente le risorse di esecuzione di un chip con una sola CPU (*single core*), ma in pratica produce miglioramenti marginali, poiché non sempre è possibile eseguire contemporaneamente istruzioni consecutive utilizzando risorse fisiche condivise. Di qui la necessità di passare alle architetture multi-core.

Microprocessori multi-core

I processori dotati di tecnologia multi-core hanno due o più unità di esecuzione (*core*) che lavorano contemporaneamente come un sistema unico e dispongono di due o più set completi di risorse di esecuzione (detti *execution core* o *computational engine*) per aumentare la potenza di elaborazione (*throughput*) del computer. Il processore multi-core si innesta su un basamento (*socket*) uguale a quello utilizzato per un chip single-core, e utilizza quindi *main board* "a una via", molto più semplici ed economiche delle main board dei server multiprocessore, mentre il sistema operativo vede ogni execution core come un processore logico distinto, con le proprie risorse di esecuzione.

I chip dual core, con due processori che lavorano come un unico sistema, sono stati la prima applicazione della tecnologia multi-core e sono i più diffusi sui PC, mentre nei server sono già largamente utilizzati chip quadri-core e stanno arrivando chip esa-core e octo-core.

Una singolarità è rappresentata dai processori Amd Phenom X3 con tre core, progettati per migliorare le performance delle applicazioni multithreaded rispetto ai processori dual-core a parità di clock, e permettono di realizzare PC capaci di gestire i workload delle applicazioni di entertainment digitale.

La tecnologia dei processori multi-core rende possibile il calcolo parallelo e può aumentare drasticamente la velocità, l'efficienza e le prestazioni dei computer senza aumentare la velocità di clock, riducendo al minimo il consumo di elettricità e la produzione di calore del sistema. I PC e server dotati di chip multi-core permettono di realizzare nuovi livelli di sicurezza e di virtualizzazione e possono migliorare la *user experience* in ambiente multitasking, soprattutto quando si eseguono in *foreground* più applicazioni in modo concorrente ad altre attività in *background*, come antivirus, comunicazioni wireless, compressione, crittografia, gestione, sincronizzazione e salvataggio dei file. I PC multi-core sono sufficientemente potenti per porsi al centro della *digital home* e gestire le applicazioni tipiche del *digital lifestyle*: digital media entertainment, contenuti multimediali "ricchi", integrazione tra PC, TV e musica digitale.

Applicazioni

La tecnologia multi-core è utile soprattutto in applicazioni e task "esigenti", come video editing, crittografia, giochi 3D, applicazioni multimediali, applicazioni scientifiche, CAD/CAM, intelligenza artificiale, riconoscimento della voce, della calligrafia, delle impronte digitali o della retina dell'occhio. Anche nel campo dei compilatori dei linguaggi di programmazione, si possono ridurre i tempi di compilazione fino al 50%. I computer multi-core possono però sfruttare pienamente le loro potenzialità solo se sono utilizzati con software multi-threaded, quindi sistemi operativi "multi-threaded" come Microsoft Windows e Linux e Solaris e applicazioni segmentabili in task concorrenti, con carichi di lavoro che possono essere eseguiti simultaneamente sui core disponibili. Un processore multi-core può eseguire thread di codice completamente separati (per esempio un thread lanciato da un'applicazione e uno appartenente al sistema operativo), o thread paralleli lanciati da un'unica applicazione. Le applicazioni multimediali beneficiano in modo particolare del parallelismo a livello di thread poiché molte operazioni possono essere eseguite in parallelo. Si prevede che nei prossimi anni gli sforzi di sviluppo si concentreranno sulle applicazioni multithread, per sfruttare sempre meglio i sistemi multi-core.

Problemi e sviluppi futuri

Attualmente, il software esistente non è ottimizzato per l'utilizzo della tecnologia multi-core. Molte applicazioni sono codificate senza pensare al consumo di risorse e non sono in grado di dividere le loro attività in flussi di lavoro trattabili separatamente in diversi core: ci vorrà ancora del tempo prima che i produttori di software riescano a creare una massa critica di applicazioni con capacità di multithreading. Inoltre, anche i sistemi operativi multi non sono ancora in grado sfruttare pienamente il potenziale della tecnologia multi-core.

Nel futuro, aumenterà sicuramente la tendenza a utilizzare la tecnologia multi-core e la prossima generazione di applicazioni software sarà multithreaded e ottimizzata per i processori multi-core. Il numero di core che può essere integrato in un unico chip aumenterà a 8, 16 e più core, mentre la dimensione fisica e il consumo di energia diminuiranno. I microprocessori multi-core sono richiesti per il *pervasive computing* poiché offrono benefici di prestazioni e produttività non ottenibili con i microprocessori single-core, e hanno un importante ruolo nella sicurezza dei PC e nei processi di virtualizzazione. Inoltre, poiché offrono maggiori prestazioni senza aumentare l'assorbimento elettrico (maggiore potenza di calcolo per watt) consentiranno di realizzare server tradizionali e *blade server* sempre più potenti e compatti. CBr