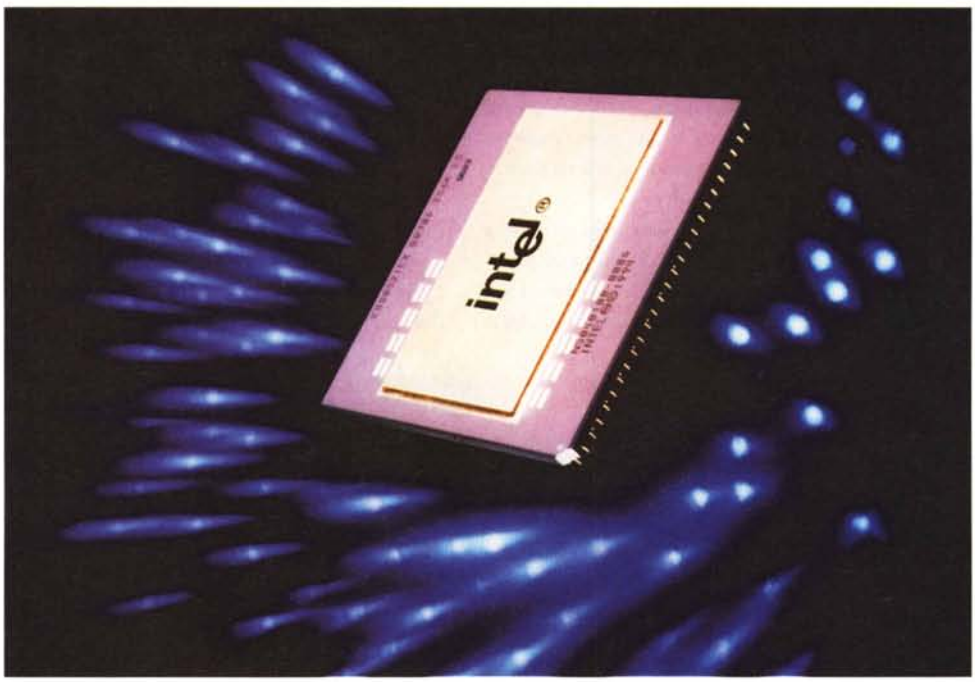
Presentato anche in Italia il successore di Pentium, una vera e propria bomba tecnologica in grado di fornire prestazioni doppie a parità di clock grazie alla sua architettura superpipeline superscalare e all'Esecuzione Dinamica che, come un abile prestigiatore, rimescola le istruzioni da eseguire alla ricerca di sequenze "salvatempo"

Intel P6

(cogito ergo sum!)

di Andrea de Prisco



Del successore di Pentium, il fantomatico Intel P6 (il nome ufficiale sarà svelato solo al momento dell'effettiva commercializzazione prevista entro la fine dell'anno... secondo me si chiamerà "Sexy"), ne abbiamo già parlato sulle pagine di MCmicrocomputer sullo scorso numero di marzo, in occasione della presentazione ufficiale della sua microarchitettura, svoltasi a San Francisco il 16 febbraio di quest'anno.

A distanza di soli tre mesi da quella fatidica data, una macchina perfettamente funzionante basata sul nuovo chip è stata mostrata alla stampa italiana, nel corso di una interessantissima presentazione tecnica nella quale sono state mostrate le caratteristiche realmente innovative di questo eccezionale "mostro" tecnologico.

Riallacciandoci al discorso lasciato in sospeso nel precedente articolo, questo mese vedremo un po' più in dettaglio la straordinaria architettura di P6, mettendo in evidenza proprio il fatto che i progettisti dei microprocessori, continuando di questo passo, non finiranno mai stupirci con le loro realizzazioni sempre più geniali.

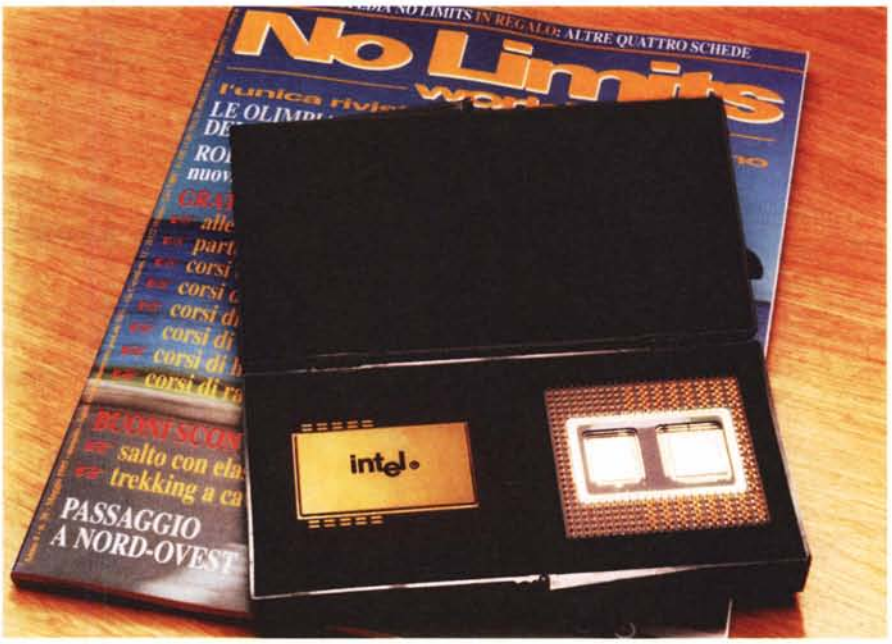
Guardando al futuro, ancora una volta, c'è da chiedersi come sarà P7, ma anche P8, P9...

Riassuntino

Per chi si fosse collegato solo in questo momento sul sesto canale (P6, per l'appunto!) riassumiamo brevemente le caratteristiche principali della nuova architettura Intel.

Innanzitutto P6 non è "un" chip ma, praticamente, "due". All'interno dello stesso contenitore ceramico, infatti, troviamo due distinti "pezzi" di silicio, il primo è il microprocessore vero e proprio (a sua volta composto di svariate unità logiche e delle immancabili cache di primo livello per istruzioni e dati) il secondo è la cache di secondo livello, attualmente disponibile in tagli da 256 o 512 Kbyte.

Microprocessore e cache sono internamente connessi da un bus ottimizzato a 64 bit ad alta velocità che permette accessi ai dati e alle istruzioni ad una velocità ben superiore di quella generalmente raggiungibile attraverso un bus e una cache esterna.
P6, con i suoi 5.5 milioni di transistor offre, senza ombra di dubbio, il più alto livello di prestazioni su architettura Intel.
Dal punto di vista logico, la più alta sofisticazione tecnologica messa a disposizione di P6 è la cosiddetta "Esecuzione Dinamica" (una combinazione di tecnologie comprendenti la predizione multipla di salto, l'analisi del flusso dati e l'esecuzione speculativa) grazie alla quale le unità di elaborazione interne sono "rifornite" ininterrottamente di dati e istruzioni. Nel loro insieme, come la stessa Intel dichiara, la combinazione di queste tre tecniche permettono al P6 di funzionare come "un'efficiente fabbrica di informazioni": sono analizzate sezioni del flusso di programma in entrata molto più ampie rispetto a quelle di qualsiasi altro processore per PC, sono allocate velocemente le risorse interne ed ottimizzati in modo intelligente i lavori che possono essere svolti in parallelo, consentendo in pratica una maggiore velocità d'elaborazione. Vediamo brevemente di cosa si tratta.
La *predizione multipla* di salto permette al microprocessore di eseguire una quantità di istruzioni maggiore senza incontrare situazioni di attesa.
L'*analisi del flusso* dei dati effettua a tempo di esecuzione un riordinamento delle istruzioni da eseguire indipendente dall'ordine stabilito del programma. Tale tecnica è parente stretta delle architetture data-flow in cui ogni istruzione viene eseguita non appena sono disponibili i dati sui quali operare, indipendentemente (se, ovvero, non sussiste dipendenza) dall'esecuzione delle altre istruzioni. I processori, per così dire, tradizionali sono invece instruction-flow: il flusso d'esecuzione è dato esclusivamente della sequenza di istruzioni di cui è formato il programma.
L'*esecuzione speculativa*, infine, consente al P6 di mantenere il proprio nucleo superscalare il più operante possibile, eseguendo quelle istruzioni che



Visione "fronte/retro" dell'Intel P6. Si notino le sue piastrine di silicio relative al processore vero e proprio e alla cache di secondo livello da 256 o 512 kbyte.

"probabilmente" saranno più necessarie... da lì a pochi colpi di clock.
Ma, come era da attendersi, P6 non si ferma qui e offrirà ulteriori interessantissime caratteristiche che semplificano la progettazione dei sistemi multiprocessor migliorando l'affidabilità dell'intero sistema.
P6 sarà disponibile nella versione a 133 MHz (con moltiplicatore di clock interno, la velocità delle board sarà, per la gioia di tutti i costruttori, ben più bassa: metà, un terzo o un quarto di quella nominale) e sarà alimentato a soli 2.9 volt. Grazie alla bassa tensione di alimentazione, il chip (nella sua totalità, microprocessore più cache di secondo livello) non consumerà più di 14 watt, mentre riguardo le performance raggiungibili si parla di oltre 200 SPECint92, pari al doppio delle prestazioni oggi raggiunte dai più veloci processori Pentium.

La "ricetta" del successo

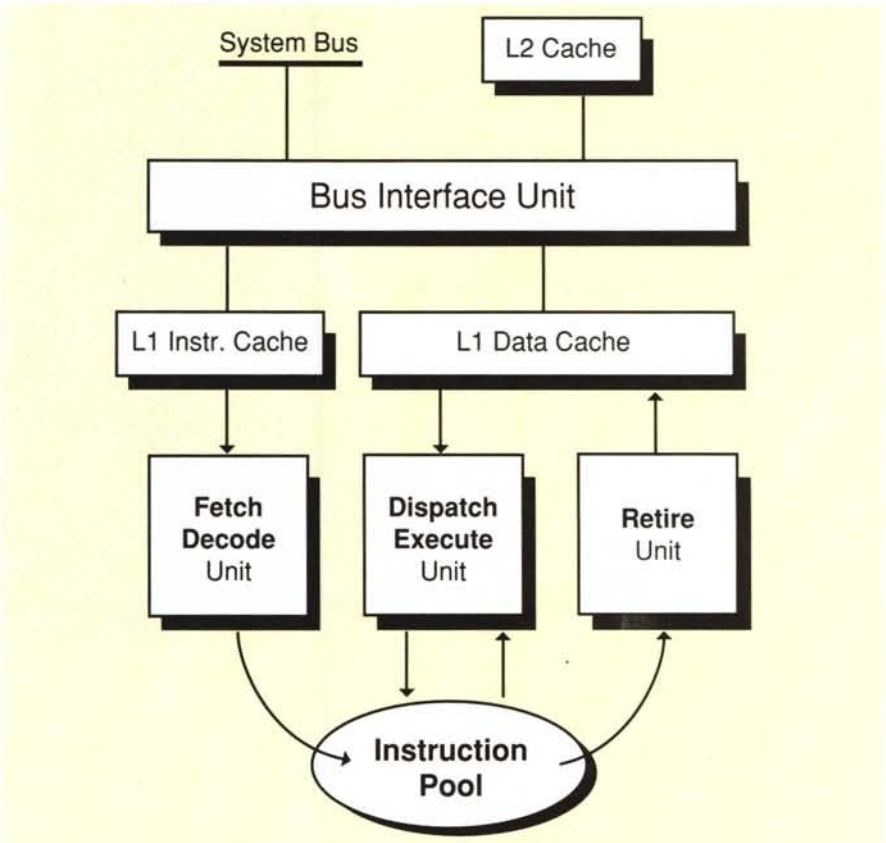
Il microprocessore Pentium (per brevità P5), è caratterizzato da una microarchitettura pipeline superscalare.
L'implementazione pipeline di P5 usa cinque stadi d'esecuzione mentre P6 utilizza un'implementazione a 12 stadi: in pratica il lavoro è ulteriormente suddiviso in un numero maggiore di stadi secondo il motto "divide et impera" (in pratica troviamo una "catena di montaggio" più lunga ma caratterizzata da step più elementari e, per questo, più veloci).
La microarchitettura superscalare di P5 è in grado di eseguire due istruzioni complete per ciclo di clock, ma è difficile superare tale valore senza rivoluzionare completamente l'approccio. La "rivoluzione" di P6 consiste nel fatto che la fase di esecuzione ha un'ampia visione del flusso di istruzioni in corso, in modo da ricercare e stabilire un diverso ordine di "lavorazione" delle stesse al fine di completare l'intero lavoro in un tempo ridotto. La funzione di un processore, infatti, non è tanto quella di eseguire singole istruzioni quanto eseguire programmi, flussi di istruzioni. Visto che non sempre è necessario eseguire nello stesso ordine imposto dal programma le singole istruzioni, quello che fa P6 (nonostante possa sembrare folle) è cercare all'interno della sezione eseguita una differente sequenza che faccia risparmiare tempo.
Facciamo un esempio "extrasettore". Immaginiamo di dover preparare una torta. Abbiamo tutti gli ingredienti (i dati) ed una ricetta (il programma) che recita

più o meno così:

1) prendere un certo numero di uova e separare l'albume dai tuorli
2) battere una parte dei tuorli con lo zucchero e la farina, aggiungendo lentamente latte caldo
3) mettere il tutto sul fuoco mescolando continuamente fino all'ottenimento di una crema densa
4) montare l'albume a neve
5) battere una parte dei tuorli con lo zucchero e la farina
6) miscelare albume montato e tuorli battuti
7) aggiungere il lievito per dolci
8) infornare il pan di Spagna
9) guarnire con la crema il pan di Spagna dopo 30 minuti di cottura.

Analizzando la ricetta, è possibile evincere che alcune fasi possono essere sovrapposte (facendoci aiutare da qualcuno possiamo eseguire simultaneamente i punti 4 e 5) alcune sono propedeutiche di altre (ad esempio non possiamo battere i tuorli, punto 2, prima di averli separati dall'albume o guarnire il pan di Spagna prima di averlo preparato e cotto). Questo, per quanto possa sembrare strano, è un tipico ragionamento da Pentium... negato ai precedenti 486 per l'assenza della doppia pipeline interna.
P6, molto più intelligente di P5, prima di iniziare l'esecuzione dà un'occhiata all'intera ricetta e si accorge che può ulteriormente risparmiare una buona decina di minuti (il tempo stimato di cottura della crema) cambiando l'ordine delle istruzioni. Ci pensa un attimo e (mentre conclude i "lavori" precedenti) formula questa nuova sequenza:

1) prendere un certo numero di uova e separare l'albume dai tuorli
2) (ex 4 & 5) montare l'albume a neve & battere una parte dei tuorli con lo zucchero e la farina
3) (ex 6) miscelare albume montato e tuorli battuti
4) (ex 7) aggiungere il lievito per dolci
5) (ex 8) infornare il pan di Spagna
6) (ex 2) battere una parte dei tuorli con lo zucchero e la farina, aggiungendo lentamente latte caldo



Schema a blocchi dell'interfacciamento con la cache L1, la cache L2 e il bus di sistema. L'Instruction Pool si occupa di generare nuove sequenze di elaborazione "salva tempo".

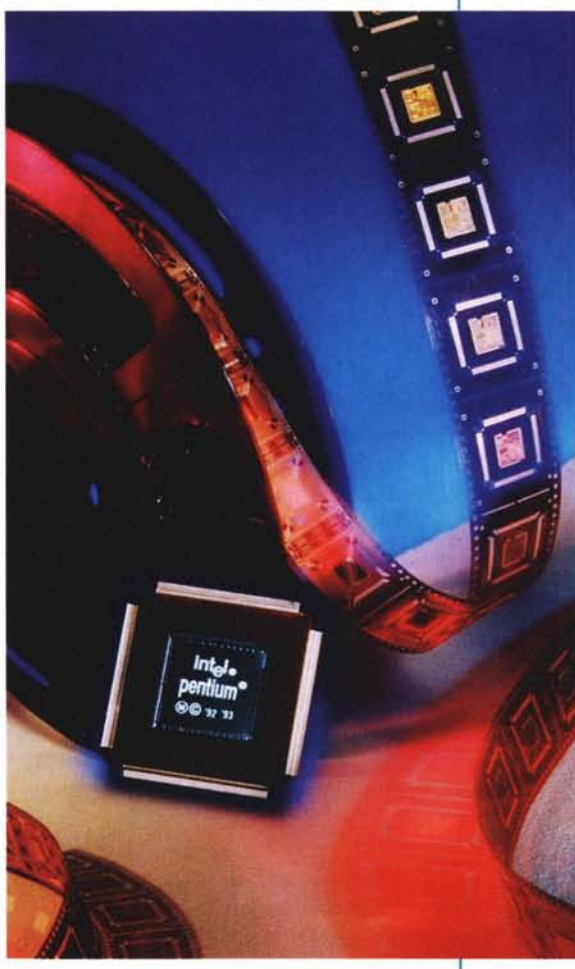
7) (ex 3) mettere il tutto sul fuoco mescolando continuamente fino all'ottenimento di una crema densa
8) guarnire con la crema il pan di Spagna dopo 30 minuti di cottura.

Preparando la crema *durante* la cottura del pan di Spagna è possibile eseguire l'intero compito (la realizzazione della torta, l'unica cosa che effettivamente ci interessa!) in un tempo minore senza alcuna differenza qualitativa o quantitativa riguardante il risultato. L'analisi effettuata da P6 consiste proprio nell'individuare quelle istruzioni che possono essere eseguite indipendentemente dalle altre anche nel caso in cui si trovino "più avanti" nel programma, sfruttando i tempi morti dei "cache miss". Se, infatti, un dato non è disponibile nella cache, inoltrata la richiesta alla memoria, P6 esegue altre istruzioni che non dipendono logicamente dall'istruzione iniziata ma non ancora completata. Non appena arriva il dato richiesto anche l'istruzione sospesa può essere completata, e così avanzeranno nell'esecuzione anche le istruzioni che dipendevano dalla prima. Ma con il vantaggio di aver eseguito "dell'altro" durante i tempi morti. Roba da non crederci!

External Bus: al servizio di Sua Maestà

Uno dei ruoli primari di un bus esterno è quello di supportare efficientemente le richieste dal parte della CPU. L'esecuzione dinamica di P6 richiede maggiori accessi al bus rispetto a quanto avviene con l'architettura Pentium e quindi il bus esterno di P6 necessita di maggiori capacità. Sul bus del Pentium, ad esempio, solo una singola richiesta può essere inoltrata per volta e il microprocessore aspetta l'esito della richiesta prima di continuare la sua esecuzione

Anche il Pentium non scherza: siamo arrivati a quota 133 MHz!



Pentium è ora disponibile anche nel formato "Tape Carrier Package" per l'assemblaggio rapido su schede elettroniche.

interna. P6, di contro, non si ferma un solo attimo e continua a processare istruzioni anche dopo un "cache miss" (l'assenza di un determinato dato richiesto all'interno della memoria cache) potendo effettuare fino a quattro successive richieste di bus prima di dover necessariamente "aspettare". Sulla documentazione relativa a P6 troviamo anche un interessante esempio chiarificatorio:

"Si immagini di essere clienti di un hotel dove è in funzione un servizio di parcheggio "Car Valet", operante come un bus Pentium. Rivolgendosi all'addetto è possibile eseguire un'operazione di lettura, "Mi serve l'auto, la prenda, per favore", oppure un'operazione di scrittura, "Per favore, parcheggi la mia auto". Se siamo l'unico cliente dell'hotel non sussistono problemi di sorta, disponendo in pratica di un "valletto" tutto per noi, sempre pronto a soddisfare ogni nostra richiesta. Ma se nell'hotel, come è altamente probabile, ci sono altri clienti che devono prendere o lasciare l'auto è facile che con un solo addetto gli stessi siano costretti a lunghe attese prima di essere effettivamente serviti. Il bus di P6 opera, in pratica, come un servizio di "Car Valet" con otto addetti che possono eseguire richieste multiple di parcheggio/prelievo d'auto e i clienti dovranno aspettare prima di fare la loro richiesta solo nel caso in cui tutti gli addetti fossero impegnati. Naturalmente deve essere definito un apposito protocollo per evitare collisioni tra le auto presso l'uscita del garage evitando entrate ed uscite (operazioni di lettura e scrittura) simultanee. Da segnalare, infine, che nel garage P6 le auto non si muovono più velocemente che in un garage Pentium, ma "solo" in maniera più efficiente: basta pensare alla differenza esistente tra la consegna simultanea di cinque autovetture, effettuata ad esempio in cinque minuti totali, e l'attesa degli stessi cinque minuti tra la consegna di ognuna di esse ad opera di un unico, affannatissimo, addetto."

P6, come già anticipato, dispone di due distinti bus dati a 64 bit. Un bus interno collega il microprocessore vero e proprio con la cache interna di secondo livello da 256 o 512 kbyte (e non si "affaccia" all'esterno del chip), un bus esterno è usato per il collegamento con la memoria di sistema, l'I/O ed altri processori. Il primo "corre" all'effettiva velocità del chip (attualmente 133 MHz), offre una larghezza di banda pari a 1 gigabyte/secondo, ed è utilizzato per tutti i "missing" della cache interna di primo livello con una percentuale di utilizzo pari al 90-95%.

Il secondo, il bus esterno, viaggia invece alle frequenze tipiche delle attuali board, 66, 50 o 33 MHz. A 66 MHz il bus esterno ha una larghezza di banda pari a 528 megabyte/secondo ed è utilizzato per tutti i "missing" della cache interna di secondo livello. In questo caso, tenuto conto della dimensione di tale cache, l'utilizzo è molto più basso (circa il 10%) consentendo la "pacifica" coesistenza di più P6 sullo stesso bus senza grossi problemi. Con quattro processori, usando come test tipici benchmark da server, il bus è utilizzato all'incirca al 60% delle sue capacità lasciando trasparire in maniera evidente l'alta scalabilità dei sistemi multiprocessor basati su P6.