

# Elettronica ed evoluzione dei Processori x86

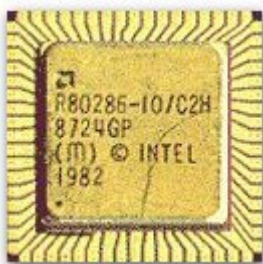
- **Processori x86 di prima generazione**

Tornando agli albori della tecnologia informatica nel 1981 Intel (Integrated Technology) introdusse il primo processore a 16 Bit per Personal Computer IBM XT: il mitico Intel 8086; Nell' 8086 sono state definite le linee guida e l'ISA (Instruction Set Architecture) dell'architettura x86 e a distanza di 20 anni, ancora oggi per garantire la compatibilità con il passato, troviamo intatti nei processori moderni elementi architettonici di questo chip capostipite della famiglia x86. La Cpu era costruita integrando nel silicio circa 29.000 transistor. La potenza di calcolo era così bassa che persino le applicazioni Dos in semplice testo a 16 colori erano lentissime ed il refresh (aggiornamento a video) dei fogli di calcolo come il Lotus 1-2-3 era così lento che durante un ricalcolo si poteva assistere alla ricostruzione della tabella dall'angolo in alto a sinistra a quello in basso a destra dello schermo. Ne venne prodotta anche una versione con bus esterno ad 8 bit detta 8088 la quale, avendo costi inferiori, fu poi quella maggiormente commercializzata.



Il Benchmark usato allora per la misura delle prestazioni era il "Norton System Information" che per la Cpu 8086 a 4.8 Mhz segnava un indice pari a "1" e saliva a 2 per l'8086 a 10 Mhz. Questo processore poteva indirizzare al massimo 1 Megabyte di Ram di cui realmente solo 640 Kbyte erano disponibili per il sistema operativo Dos.

- **Processori di seconda generazione**



L'introduzione dell'80286 nel 1982 significò un vero salto da un punto di vista tecnologico; con i suoi 134.000 transistor e frequenze tra 6 e 20 Mhz otteneva un indice Norton tra 8 e 21. Grazie alla sua capacità di gestire 16 Mbyte di Ram rese possibile l'apparire delle prime arcaiche interfacce grafiche (Windows 2.0 e 2.1). I 16 Megabyte di memoria erano però gestiti a blocchi di 64 Kbyte in quanto trattandosi di un processore a 16 bit la sua capacità di indirizzamento era limitata a soli  $2^{16}=65535$  byte per volta. Il 286, come tutti i processori Intel successivi, disponeva di una perfetta compatibilità all'indietro con i programmi Dos scritti per l'8086 mentre i programmi che volessero

usare la memoria al di sopra del megabyte dovevano accedervi con una modalità "protetta" che rendeva disponibile tramite un driver Xms (Extended memory specification) la memoria da 2 a 16 Mbyte.

## • Processori di terza generazione

Il primo 80386 Dx fu realizzato nell'Ottobre 1985 integrando 275.000 transistor, partito dai 16 Mhz arrivò fino a 40 Mhz con un indice di prestazioni Norton Si da 22 a 43. Con il suo bus interno a 32 Bit poteva indirizzare una maggiore quantità di dati e gestire una quantità di memoria Ram fino a 4 Gigabyte ( $2^{32}=4.294.967.296$  byte) contro il massimo limite di 16 megabyte dei precedenti processori Intel a 16 Bit. Inoltre con l'introduzione di questo processore è sparito il problema della segmentazione della memoria a blocchi da 64 Kbyte tipica del 286. La sigla Dx sta per Double word eXternal ed indica la capacità del processore di gestire due word (parole) di  $16+16=32$  Bit. External sta a significare che il processore



comunica verso l'esterno, ossia verso il bus di memoria della scheda madre, sempre a 32 Bit. Intel produsse in seguito anche una versione a basso costo 80386 Sx (Single word eXternal) con bus interno a 32 bit ed esterno a 16 bit. La potenza di calcolo del 386 era divenuta sufficiente a gestire un vero sistema operativo grafico di tipo Gui (Graphical User Interface) e ciò ha permesso la definitiva affermazione di Windows 3.0 e poi 3.1. In realtà queste versioni di Windows continuavano ad usare la modalità di indirizzamento della memoria segmentata a 16 Bit e bisognerà attendere Windows 95 per vedere in opera i primi software a 32 Bit. Questo processore segna anche l'entrata in campo della concorrenza di Intel: Amd

(Advanced Micro Devices) aveva appena prodotto il suo primo "clone" 386 a 40 Mhz e da qui cominciò la battaglia a suon di denunce e carte bollate tra le due grandi aziende costruttrici di processori.

### La memoria Cache

Il 386 Dx segna anche l'introduzione della tecnologia di caching della memoria. Si vide che il costoso (per l'epoca) 386 a 33 Mhz, in condizioni standard e con l'uso di comune memoria Ram dinamica (Dram Fastpage), non risultava affatto più veloce del 386 a 25 Mhz. La lentezza della memoria Ram da 80 ns (nanosecondi) a cui il processore accedeva per scrivere e rileggere dati fungeva da collo di bottiglia strozzando le prestazioni. Si pensò così di saldare sulla scheda madre un paio di chip da 32-64 Kbyte di veloce memoria Sram (Static Ram - Ram Statica) da 20 ns per velocizzare la trasmissione dati tra il processore e la memoria Ram di sistema (vedi figura).



Il meccanismo è basato sul principio che alcuni dati, appena impiegati, possano essere richiesti di nuovo per la successiva elaborazione. Se gli stessi vengono quindi memorizzati in un'area di memoria ad accesso ultrarapido il processore può avere immediato accesso agli stessi senza stare a richiederli di nuovo alla lenta memoria Ram. Per fare un esempio pratico prendiamo una Cpu a 386 a 40 Mhz funzionante su una scheda madre dotata di Ram dinamica da 80 ns; a 40 Mhz il 386 impiega 20 ns per completare un ciclo di elaborazione, ogni accesso alla memoria centrale deve durare quindi almeno quattro cicli di clock ( $80/20=4$ ). Abbiamo detto "almeno" perché poi la Ram dinamica ha anche dei cicli di attesa (Wait States) che possono far lievitare a 6 i cicli di clock nei quali il processore rimane in stato di Idle (ozio) ossia a "girarsi i pollici" in attesa che la memoria Ram fornisca i dati richiesti. Ciò accade perché la Ram dinamica Dram è costruita in modo tale da trattenere le informazioni in essa memorizzate solo per un brevissimo lasso di tempo e quindi richiede un continuo rinnovo (refresh) del proprio contenuto, sia che

le informazioni (i bit di dati) in essa presenti vengano aggiornati o meno. La necessità del refresh della memoria DRam dipende dal fatto che i singoli bit sono registrati per mezzo di transistor in celle che mantengono, a mo' di condensatori, una carica elettrica. Se la cella è

carica il Bit vale 1, se è scarica vale 0. Esiste a tale scopo un apposito circuito che si occupa di effettuare il refresh delle celle di memoria ogni x cicli di clock della Cpu. Possiamo quindi immaginare la memoria Ram dinamica di un computer come una smisurata griglia di celle atte a contenere i dati che di volta in volta il processore richiede. La RAM statica invece può conservare meglio i dati più poiché, essendo le sue celle in grado di trattenere a lungo la carica elettrica, viene meno il bisogno di effettuare continui refresh. Frapporre quindi una piccola quantità di memoria cache Sram, ossia una memoria di transito veloce da 20 ns, tra il processore e la memoria Ram dinamica di sistema può far sì che il 386 a 40 Mhz possa accedere ai dati in un sol ciclo di clock aumentando di fatto le prestazioni nell'accesso alla memoria del 400-600%. Questa piccola e costosissima memoria Sram, è stata appunto definita memoria cache e contiene i dati più prossimi alle unità di esecuzione del processore. Abbiamo voluto a lungo sottolineare questo aspetto poiché, come è facile immaginare, con l'aumentare delle frequenze in Mhz dei processori si è verificò (e continua tuttoggi a verificarsi) il problema dei cicli di latenza delle memorie Ram che costituisce oggi il maggior impedimento alle prestazioni degli attuali processori. Vedremo nel seguito come di recente questo problema è stato affrontato.

### **Come ragiona un Processore?**

Un microprocessore è uno speciale circuito integrato che, facendo uso di logica digitale, processa una serie di bit che contengono informazioni (numeri) sotto il controllo di altre serie di bit che compongono invece le istruzioni da applicare su tali numeri. Le operazioni vengono svolte in particolari aree di memoria dette registri all'interno del processore. Facciamo un esempio di un semplice calcolo: il processore per prima cosa carica uno dei numeri in uno dei suoi registri, poi ne carica un altro in un secondo registro. Quindi legge l'istruzione di programma che dice al chip quale particolare operazione dev'essere svolta (ad esempio una somma). L'istruzione attiva un altro minuscolo programma che si trova all'interno di una speciale unità di decodifica che impone ai circuiti del chip di calcolare il risultato dell'operazione e di porlo in un altro registro. Calcolato il risultato una successiva istruzione permette l'uscita della risposta verso la memoria Ram e di qui verso una unità di Output (il monitor od una stampante). In pratica quindi un processore "ragiona" in termini seriali, ossia esegue un determinato numero di istruzioni una di seguito all'altra richiamando dati dalla memoria Ram di sistema ed emettendo dati ad operazione effettuata. Gli attuali computer sono pertanto in grado di essere molto veloci nel calcolare dati ma essendo seriali non possono elaborare istruzioni diverse da quelle introdotte in Input in partenza. Al contrario il cervello umano, per fare un esempio, è un sistema di elaborazione a parallelismo massiccio (una rete di 10 miliardi di neuroni), poco efficiente in termini puramente computazionali ma capace di elaborare dati eterogenei provenienti contemporaneamente da più fonti.

### **CPU, ALU ed FPU**

I processori di classe x86, possiedono due unità di calcolo distinte: la Alu e la Fpu. La Alu (Arithmetic Logic Unit - Unità Aritmetico Logica) serve per processare i numeri interi ossia i numeri naturali positivi 1, 2, 3, ecc. e negativi -1, -2, -3 ecc. nonché operazioni di logica booleana. La maggior parte dei software di videoscrittura (Word), database (Access) e di grafica 2D usa la unità Alu del processore. La unità Fpu (Floating Point Unit - Unità di calcolo in virgola mobile) serve invece per processare i numeri con virgola ossia quelli razionali frazionari ed irrazionali (3.14 ad esempio) e quindi per il calcolo di divisioni, radici, funzioni trigonometriche, logaritmi ecc.. Questo tipo di calcolo è usato da applicativi software di ingegneria come Autocad, di calcolo scientifico e, in parte, dai fogli di calcolo come Excel. Negli ultimi tre anni sempre più applicativi multimediali vanno ad usare queste istruzioni ed in particolare i sempre più diffusi videogiochi 3D (Quake I-II-III, Formula 1 Gp, ecc.) e le applicazioni di streaming video quali Flaskmpeg, Adobe premiere ecc.

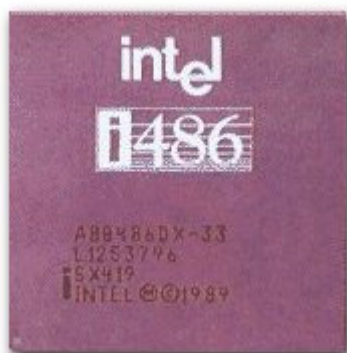


La unità Fpu è detta anche "Coproprocessore matematico" in quanto sui processori 8086, 80286 e 80386 la Fpu era situata su un chip esterno detto 8087, 80287 ed 80387. E' per questo che le istruzioni classiche in virgola mobile dei processori Intel ed Amd sono dette anche x87. Nella figura potete osservare un chip coprocessore matematico 387 a 33 Mhz prodotto da Cyrix che andava inserito in uno zoccolo a parte vicino a quello del processore 386 Dx 33 Mhz di cui potenziava le capacità di elaborazione dei numeri con virgola di circa 20 volte lavorando in parallelo

ad esso. L'insieme della unità Alu e della Fpu forma la Cpu (Central Processing Unit) ed è per questo che Cpu e processore significano la stessa cosa.

## • Processori di quarta generazione: l'80486 e la Cache integrata

Con l'introduzione del primo Intel 80486 un piccola porzione di cache venne inserita all'interno dei microcircuiti nel nucleo (core) del processore, il quantitativo era limitato a soli 8 Kbyte ma, essendo la cache integrata il doppio più veloce di quella esterna su scheda madre, gli 8 Kb erano sufficienti a far ottenere un raddoppio netto delle prestazioni rispetto al 386. Grazie ad un nuovo algoritmo questa piccola cache integrata non solo immagazzina i dati impiegati più di recente come le cache Sram su scheda madre ma anticipa anche gli accessi del processore importando una certa quantità di dati dalla memoria di sistema anche quando gli stessi non sono al momento richiesti dal software. Questa funzione, detta Read-Ahead (lettura anticipata), rende disponibili al processore anche una certa quantità di dati che, con elevata probabilità, verranno poi effettivamente richiesti dall'applicativo.



La maggiore efficienza computazionale del 486 derivava quindi da tre fattori fondamentali:

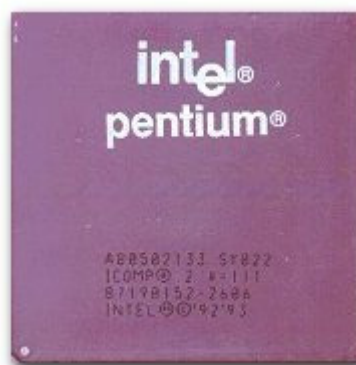
- Maggiore integrazione dei microcircuiti (a 1 micron) che ha permesso di elevarne la frequenza operativa in Mhz.
- Memoria cache integrata da 8 Kbyte a quattro vie con algoritmo Read-Ahead.
- Integrazione nel nucleo dell'elettronica del coprocessore matematico 80387 che era invece prima disponibile solo su un chip esterno.

Questi tre aspetti costruttivi del 486 hanno fatto sì che il nucleo di questo processore arrivasse ad integrare ben 1.200.000 transistor (300.000 erano per l'80387) con un raddoppio delle prestazioni della unità Alu e la triplicazione della potenza di calcolo della unità Fpu. Tale gap prestazionale nel calcolo dei numeri in virgola mobile rispetto alla accoppiata 386+387 è dovuta in parte alla memoria di transito ad alta velocità da 8 Kbyte integrata ma soprattutto alla riduzione delle distanze circuitali tra i vari elementi. In elettronica il segnale deve seguire percorsi più brevi possibile per limitare al massimo sia le interferenze elettromagnetiche che le dispersioni del segnale stesso. Raggruppare i tre componenti ed integrarli sullo stesso quadratino di silicio ha quindi ridotto la lunghezza delle connessioni all'ordine dei millesimi di millimetro contro gli svariati centimetri di rame che è necessario stendere su un circuito stampato per unire i tre componenti (Alu, Fpu e Cache) a sé stanti. I primi esemplari del 486 presentati nel 1989 funzionavano con una frequenza di 25 MHz e sono in seguito passati ai 33 MHz. Il 486 Sx era invece una versione economica privata del coprocessore matematico 80387 integrato. Quando Intel ha però introdotto il modello 486 Dx a 50 Mhz i produttori di schede madri fecero notare che una frequenza così elevata (per l'epoca) poteva introdurre disturbi di segnale e correnti parassite sulle piste delle schede madri. Di conseguenza intel produsse il processore 486 Dx2 il quale per via di un moltiplicatore interno 2x (leggi "2 per") poteva andare a 66 Mhz pur funzionando su un bus di sistema a 33 Mhz. In seguito venne introdotto il

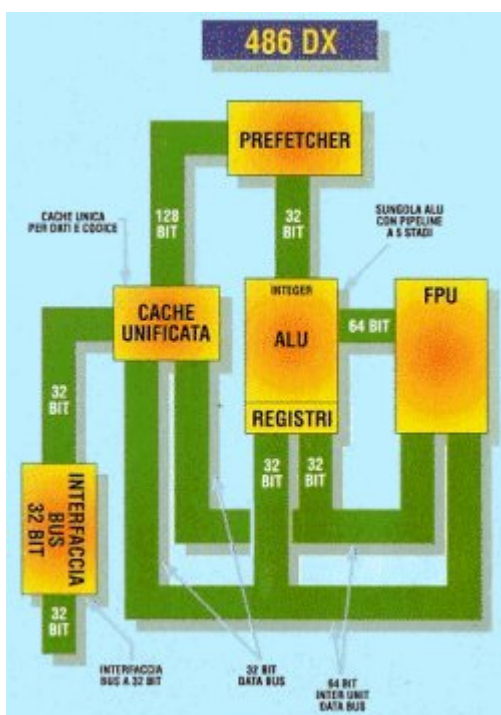
486 Dx4 a 100 Mhz con moltiplicatore interno 3x ( $33.3 \times 3 = 100$ ). La tecnica del moltiplicatore interno è stata portata all'estremo negli attuali processori e nel Pentium 4 in particolare. Avrete dunque capito che ad impostare la frequenza in Mhz del processore è un apposito circuito presente sulla scheda madre. Questo circuito detto "clock generator" è un oscillatore al quarzo che genera la frequenza di bus di sistema, frequenza dalla quale poi si ricavano, tramite moltiplicatori o divisori quella del processore e di tutti gli altri componenti (Memoria Ram, bus Pci e Agp ecc.). L'indice di prestazioni Norton Si variava dai 54 per il 486 a 25 Mhz ai 290 per il 486 a 133 Mhz di Amd.

## • Processori di quinta generazione: la rivoluzione Petium

Nel marzo del 1993 Intel introdusse il primo Pentium con frequenza di 60 Mhz, Intel non lo chiamò, come sembrava scontato, "586" al fine di aggirare la legge americana che non consente di riservarsi il copyright di un numero. La dicitura "Pentium" diveniva quindi un diritto d'autore di Intel e nessuna altra società avrebbe potuto usare in seguito quel nome.



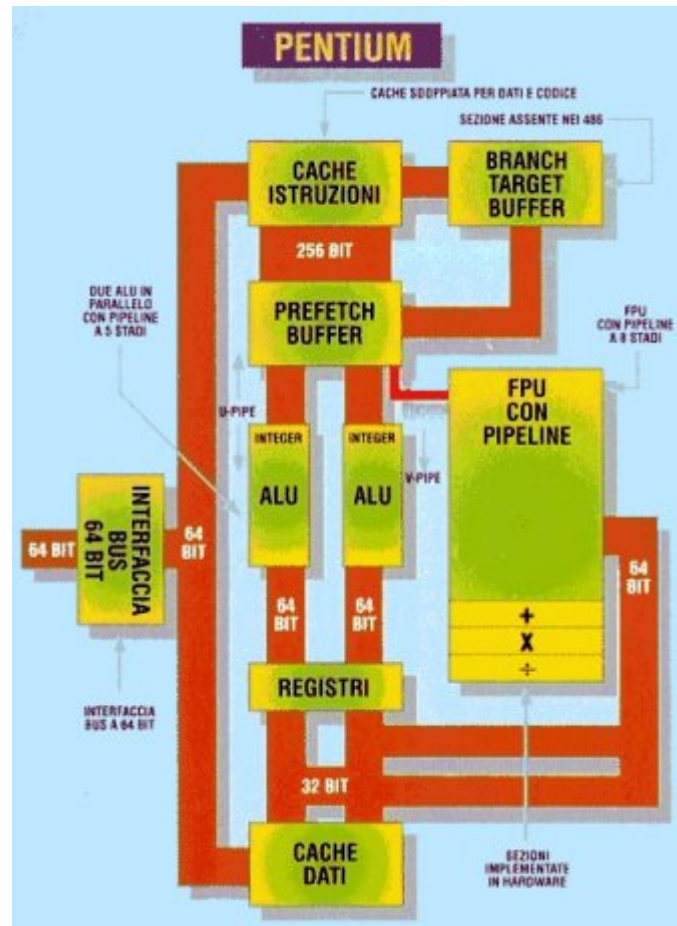
Il Pentium venne realizzato integrando nel silicio 3.100.000 transistor e rispetto ai 486 proponeva tecniche avanzate di gestione interna dei dati come una Branch Prediction Unit (unità di predizione dei salti) e tecniche di parallelismo nell'elaborazione del codice che stanno alla base di tutti i più recenti processori. Capire come funziona un Pentium è dunque importante per comprendere le strategie tecnologiche che hanno permesso di sviluppare processori negli ultimi 10 anni. Analizzeremo pertanto in profondità le caratteristiche di questa Cpu mettendo a confronto la sua struttura con quella molto più semplice del 486.



Nello schema che vedete di lato si può osservare, nella struttura logica del nucleo di un 486, la presenza di una unità aritmetico logica Alu per il calcolo degli interi ed una Fpu 80387 collegate tra loro da un bus a 64 Bit. La cache da 8 Kbyte è unificata per i dati e le istruzioni ed è collegata ad entrambe le unità di calcolo Alu ed Fpu da un bus a 32 Bit. La unità prefetcher è adibita a reperire blocchi di istruzioni dalla memoria Ram e spostarli nella cache con la quale comunica con un bus a 128 Bit. L'interfaccia di comunicazione verso il bus della memoria è infine a 32 bit.

Nel Pentium (nella figura visibile di seguito) la prima cosa che risalta è la presenza di due distinte cache da 8 Kbyte ciascuna. La prima cache è destinata alle istruzioni (codice di programma) e la seconda ai dati a cui tali istruzioni vanno applicate. Il vantaggio rispetto alla cache del 486 sta nella riduzione dei conflitti tipici di una cache unificata. Oltre alla cache sono state raddoppiate anche le unità di elaborazione Alu; in

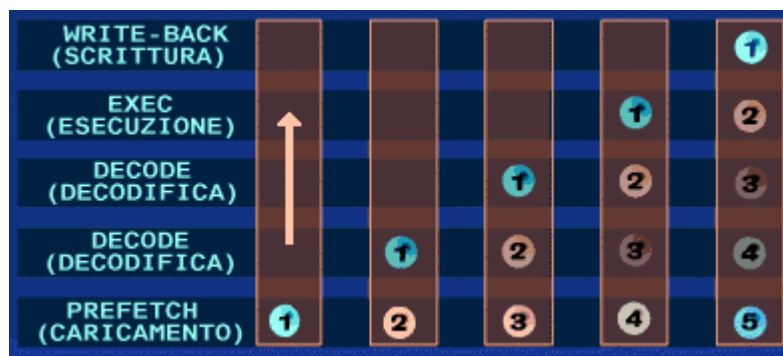
pratica nel Pentium è come se operassero due Alu 486 in parallelo. Dopo aver ricevuto e parzialmente decodificato una istruzione il Pentium stabilisce se esiste la possibilità di eseguire in parallelo l'istruzione stessa insieme a quella successiva. Non tutte le istruzioni possono essere parallelizzate ma in quelle ove è possibile applicare questa tecnica l'elaborazione richiederà tempi dimezzati. Un processore che riesce a smistare le istruzioni fra più pipeline e a parallelizzare la loro esecuzione è definito Superscalare ed il Pentium è il primo processore x86 dotato di questa proprietà. In particolare le istruzioni vengono prelevate, decodificate ed eseguite nelle due Alu che vedete in figura. In condizioni ottimali il Pentium può pertanto elaborare due istruzioni (operanti sui numeri interi) per ciclo di clock.



Una ulteriore innovazione nel Pentium è la introduzione della Branch Prediction Unit. Trattasi di una piccola zona di memoria cache di appena mezzo Kbyte strutturata in forma tabellare ed adibita a contenere una History (cronistoria) degli indirizzi a cui il software sta eseguendo salti condizionati (tipo If-Then-Else). In base a ciò, sfruttando il fatto che molti programmi eseguono più volte la stessa routine di salto, si cerca di prevedere a che punto avverrà il salto successivo. In base alla previsione il dato viene caricato nel buffer e se il salto avverrà proprio in quella posizione l'elaborazione ne sarà ampiamente agevolata.

Questa feature da sola è in grado di fornire al Pentium il 20-25% di prestazioni in più rispetto al 486 ove i salti nell'esecuzione del codice non sono precognizzati. Domanda; cosa è un salto? Il codice dei programmi è in gran parte di tipo sequenziale ma di tanto in tanto (in media una volta su otto) il programma richiama routine di codice diverse poste a monte (GoTo all'indietro) o, più di rado, a valle (GoTo in avanti) del codice in esecuzione corrente interrompendo la sequenzialità delle istruzioni. Tutto questo non è ben gradito ad un processore che si trova così di fronte ad una condizione di Branch (salto) ed è per questo che è stata ideata la Branch Prediction Unit. E se la previsione non va a buon fine cosa accade? Una previsione errata comporta lo svuotamento completo delle Pipeline del processore. Con il termine Pipeline (canalizzazione) si usa indicare la catena di montaggio delle istruzioni interna

al processore; nel 486 ne abbiamo una e nel Pentium, come abbiamo visto, ne abbiamo due che lavorano in parallelo. Tali pipeline sono divise in cinque diversi stadi: incanalando le istruzioni in una Pipeline a più stadi e suddividendo il processo elaborativo tra gli stadi stessi la Cpu può cominciare ad elaborare il primo stadio di una successiva istruzione mentre quella in fase di elaborazione corrente è appena passata allo stadio successivo. Come si può osservare nella schema mentre la istruzione (1) ha passato il primo stadio la (2) inizia ad essere elaborata. Quando la (2) è salita al secondo stadio la (1) è salita al terzo e la (3) è in ingresso ecc. Il vantaggio (teorico) è dunque di poter elaborare una istruzione per ogni singolo ciclo di clock per ogni pipeline. Abbiamo detto "teorico" perchè una condizione di Branch nel 486 così come una previsione di Branch errata nel Pentium crea un punto di ingorgo nella canalizzazione che costringe ad eliminare tutte le istruzioni precedenti e le successive a tale punto. Questa condizione è nota anche come Stallo della pipeline.

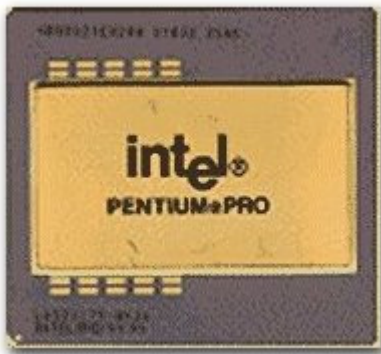


La potenza della unità coprocessore matematico (Fpu) del Pentium è praticamente quadrupla rispetto a quella del 486 e questo lo si deve al fatto che sono state introdotte nel nucleo del processore tre nuove sezioni che eseguono direttamente in hardware calcoli di addizione, moltiplicazione e divisione. Da un punto di vista dell'elettronica le primissime versioni di Pentium a 60 e 66 Mhz erano integrate a 0.8 micron con alimentazione a 5 volt per poi passare ai 0.6 micron nelle successive release da 75 a 166 Mhz con alimentazione a 3.3 Volt. La tecnica di integrazione usata era la nuova Bimos che fa uso di transistori bipolari in quei punti della microcircuiteria ove necessitano tempi di risposta rapida e di transistor Cmos (complementary metal oxide semiconductor) per la restante parte. Questi ultimi fanno uso di strati di isolante sul silicio e sulle micropiste in alluminio per ridurre le correnti e le dispersioni elettromagnetiche contenendo i consumi e la potenza dissipata dai circuiti. L'indice Norton Si variava dai 190 per il modello a 60 Mhz a 638 per il modello a 200 Mhz.

### **Pentium MMXe le istruzioni multimediali**

E' stato introdotto nel Gennaio 1997 ed è stato l'ultimo processore della quinta generazione. Sfruttando una integrazione a 0.35 micron ha potuto essere alimentato a soli 2.8 volt crescendo in frequenza dai 166 ai 233 Mhz (266 Mhz nella versione per Notebook) con un indice Norton Si di 750. Il bus rimaneva a 64bit 66MHz. Era più veloce del Pentium classico del 8-10% a parità di frequenza in quanto disponeva di una cache L1 raddoppiata a 32 Kbyte (16+16) che ha fatto lievitare il numero dei transistor integrati a 4.5 milioni. La vera novità sta però nella introduzione di 57 nuove istruzioni nel codice base x86. Queste istruzioni, dette Mmx (MultiMedia Extension) sono state il primo tentativo di estendere il codice base x86 ed adattarlo alle nuove applicazioni multimediali di grafica 2D e (in parte) 3D, streaming video, audio, riconoscimento e sintesi vocale. Trattasi di istruzioni di tipo Simd (Single Instructions Multiple Data) ciascuna delle quali può operare su diversi blocchi di dati sfruttando le unità di elaborazione parallele interne al processore. Le istruzioni eseguite dai software multimediali infatti ben si prestano ad essere parallelizzate in quanto sono costituite per lo più da loop (cicli) ripetitivi ed operano spesso sugli stessi gruppi di dati. Le istruzioni MMX operano su 64bit alla volta, configurabili secondo l'applicazione specifica come 8 word da 8bit, 4 word da 16bit, o 2 word da 32bit. Tutti i successivi processori x86 hanno poi adottato queste istruzioni ma il loro sfruttamento reale da parte dei programmatori di applicativi software ha tardato un paio di anni prima di venire implementato.

- **Processori di sesta generazione: Pentium MMX, II e III**



Per aumentare la potenza di elaborazione di un processore si può agire per due vie: aumentare la sua frequenza operativa (Clock) in Mhz oppure migliorare le sue capacità IPC (Instructions per Clock) ossia la possibilità di elaborare, parallelizzandole, più istruzioni per ciclo di clock. La parola clock in inglese letterale significa orologio ma in questo contesto è da intendersi come "temporizzatore"; parliamo cioè di un apposito circuito oscillante al quarzo che genera una determinata frequenza sulla quale vengono temporizzati diversi eventi logici. Il Pentium Pro ha rappresentato un vero salto generazionale. Costituito da 5.5 milioni di transistor integrati a 0.6 micron questo processore superscalare

implementa delle nuove tecniche di elaborazione dati che possiamo riassumere in:

### **Cache L2**

Integrata nel package: oltre alla cache di primo livello da 32 Kbyte anche la cache di secondo livello da 256 Kbyte è stata integrata nel chip per fornire più rapidamente i dati alle unità di esecuzione. In realtà la cache non è integrata sullo stesso pezzo di silicio (die) ma su una porzione separata che condivide con il nucleo principale lo stesso package ed un canale di comunicazione preferenziale. Questo rendeva il Pentium Pro costosissimo da produrre ma tale direzione, abbandonata con il Pentium II sarà reintrodotta in seguito nel Celeron e da lì in tutti i processori successivi.

### **Superpipeline**

E' stata aumentata a 14 la profondità delle pipeline di esecuzione delle istruzioni, più stadi di preparazione intermedia delle operazioni permettono di mantenere le unità di elaborazione sempre occupate e consentono di accrescere la frequenza operativa in Mhz del processore.

### **Superscalarità spinta**

Sono state portati a tre i canali di elaborazione parallela delle istruzioni contro i due del Pentium. Possiamo dire, con buona approssimazione, che il Pentium Pro implementa al suo interno tre 486 operanti in parallelo.

### **Esecuzione fuori ordine (Out of order)**

Nel Pentium, come abbiamo visto, era possibile l'esecuzione contemporanea di due istruzioni utilizzando due pipeline separate; l'esecuzione era legata alla sequenza definita dal programma, perciò ogni volta che un'operazione non poteva essere eseguita subito a causa di un stallo, entrambe le pipeline restavano ferme. Nel Pentium Pro invece le operazioni x86 vengono convertite in istruzioni micro-ops (micro-operazioni) con una tecnica che ricorda i processori Risc. Attraverso questo passaggio si eliminano molte delle limitazioni tipiche del set di istruzioni x86, cioè la codifica irregolare delle istruzioni e le operazioni sugli interi che richiedono il passaggio di dati dai registri interni alla memoria. Le micro-ops vengono quindi passate a un motore di esecuzione capace di eseguirle fuori ordine, modificandone la sequenza così da mandare in esecuzione quelle pronte e lasciare in attesa quelle che non sono. Con ciò se una Pipeline nel Pentium Pro va in stallo le altre due possono continuare ad operare senza essere svuotate. La sequenza delle istruzioni viene infine riordinata da una apposita sezione hardware detta Reorder Buffer alla fine della elaborazione.

## Esecuzione speculativa

Nel Pentium Pro le funzioni di predizione dei salti sono state potenziate ed oltre alla unità di predizione dei salti è presente una elaborazione speculativa. Essa consiste nell'eseguire istruzioni che si trovano al di là di un'istruzione di salto prima che quest'ultima sia stata eseguita e che quindi si sappia con certezza che esito avrà la diramazione. Il processore non può naturalmente aggiornare i registri interni o la memoria centrale con i risultati "speculativi" ma deve aspettare il responso della unità di branch. In caso di errata predizione del salto, il processore deve essere in grado di ritornare sui propri passi azzerando tutte le operazioni già eseguite che si riferiscono a istruzioni collocate oltre il punto di salto. Nel caso la speculazione risultasse poi sbagliata le istruzioni speculative vengono cancellate prima che giungano alla fase di termine. E' un po' il meccanismo logico che usano gli speculatori di borsa che vendono azioni non appena una società comincia ad andare male pur non avendo dati certi su un suo effettivo crollo ne su una sua possibile ripresa.

Per quanto possa sembrare strano dopo l'architettura del Pentium Pro, che è stata la base di realizzazione di tutti i processori successivi, non ci sono state grandi mutazioni tecniche strutturali nel miglioramento delle prestazioni delle Cpu Intel atte ad aumentare il fattore Ipc.



Il nucleo del Pentium II deriva direttamente dell'architettura "P6" del Pentium Pro. Addirittura si potrebbe parlare quasi di una semplificazione (es: riduzione delle pipeline da 14 a 10 stadi, finestra per l'esecuzione fuori ordine e speculativa più piccola, etc..) con l'obiettivo di concentrarsi più sull' aumento della frequenza operativa che non sull'aumento del fattore Ipc (Instructions per Clock). Con il Pentium II si cambia anche formato, dal vecchio processore su zoccolo (socket) Intel passa ad un nuovo formato proprietario denominato Slot I consistente in un connettore a pettine nel quale alloggiare una "cartuccia" processore contenente il chip vero e proprio e un certo quantitativo di cache L2 ad alta velocità (512KB a metà frequenza nei modelli normali, 1MB o 2MB a piena frequenza nei modelli Xeon dedicati ai server). Lo slot I sarà disponibile prima a 66MHz e successivamente a 100MHz per meglio supportare i processori con frequenze di circa 400MHz.

Dopo il Pentium II è la volta del debutto del Celeron. Intel si rende conto che l'assemblamento della cartuccia e della cache di secondo livello rendono il prodotto troppo costoso e difficile da offrire alla fascia bassa del mercato per cui prima immette sul mercato una versione del Pentium II priva della cache di secondo livello, e poi viste le scarse prestazioni di questa soluzione, mette a punto un processore dotato di cache L2 ridotta a 128K ma direttamente integrata nel die di silicio. Il Celeron debutta a 300MHz con bus a 66MHz e per molto tempo sarà il cavallo di battaglia di Intel nel settore entry-level vendendo moltissimi pezzi e sbaragliando la concorrenza di AMD. Con il Celeron debutta anche la tecnologia di integrazione a 0.25 micron con tensione di core di 2V, utilizzata anche nei Pentium II dai 350MHz in poi.

Al Pentium II succede il Pentium III, invero estremamente simile al suo predecessore. Il PIII debutta a 450MHz, utilizza ancora lo Slot I ed ha un FSB (Front Side Bus) a 100MHz. Internamente, a parte piccole ottimizzazioni, la innovazione più importante è costituita dall'introduzione delle estensioni SSE (note anche come KNI). Le SSE rappresentano per i numeri in virgola mobile quello che le MMX rappresentano per i numeri in virgola fissa. In pratica viene potenziata l'unità floating point per poter gestire operazioni di tipo SIMD (Single Instruction Multiple Data). L'unità manipola dati ampi 128bit e configurabili come 4 numeri floating point a singola precisione (32bit) o come 2 numeri a doppia precisione (64bit). Le SSE risultano utili nella manipolazione dei contenuti multimediali, nella decompressione dei filmati MPEG2 (DVD) e nell'elaborazione della geometria 3D (T&L).



L'ultima evoluzione dell' architettura P6 porta al Pentium III Coppermine. Il Core Coppermine integra nel die 256KB di cache L2 a piena frequenza, bassa latenza ed elevata banda (256bit di collegamento con il core). Con il Coppermine si torna anche al socket (socket 370), abbandonando il costoso Slot I, e si raggiungono i 133MHz di FSB. Intel coglie anche l'occasione per inaugurare la nuova tecnologia a 0.18 micron che porterà il Coppermine dai 600MHz del debutto alle soglie dei 1100MHz, frequenza alla quale la tecnologia P6 comincia a mostrare tutti i suoi limiti costringendo Intel ad un lungo stallo nell'avanzamento tecnologico, stallo che durerà fino alla presentazione del Pentium 4 e del Pentium III Tualatin.

Mentre il Pentium 4 rappresenta di fatto il passaggio per Intel dalla sesta alla settima generazione, il Pentium III Tualatin non è altro che l'ennesima ottimizzazione della collaudata architettura P6 tesa a servire il mercato dei portatili e dei server in attesa della definitiva stabilizzazione della piattaforma Pentium 4. Il core Tualatin annovera una cache L2 integrata da ben 512KByte e una tecnologia di 0.13 micron, per il resto risulta identico al Pentium III. Nuova veste anche per il Celeron il cui bus passa a 100MHz e che nelle ultime versioni eredita la tecnologia Tualatin a 0.13 micron che permette l'integrazione di 256KByte di cache.

AMD è praticamente da sempre attiva nel mondo dell'elettronica, in particolare nell'ambito delle memorie, delle logiche programmabili e, ovviamente, dei processori. Benchè adesso sia acerrima avversaria di Intel, un tempo le cose tra i due colossi non stavano in questi termini, ma anzi c'era collaborazione e (addirittura) amicizia, ma partiamo dall'inizio...

Advanced Micro Device nasce nel 1969, nel '76 inizia la partnership con Intel con la quale firma un accordo per lo sfruttamento dei microcodici degli allora neonati processori Intel, nel 1982 AMD e Intel rafforzano la collaborazione firmando un accordo decennale in cui si impegnano a lavorare congiuntamente sulla piattaforma x86. Nel '87 arriva però la rottura, AMD accusa Intel di non aver rispettato i patti e vince la causa nel 1992 anno in cui comincia a sviluppare autonomamente un processore alternativo a quello Intel.

Ai tempi del 486, quando ancora erano in corso gli accordi di collaborazione con Intel, AMD e Cyrix (altro produttore di processori e coprocessori x86) detenevano addirittura quote di mercato del 30%. I prodotti AMD erano in pratica "cloni" dei processori Intel e Intel sopportava la situazione in nome di un principio noto nel mondo dell'elettronica come "Second Source"; in pratica per l'affermazione di un chip è necessario che esistano sul mercato almeno due fornitori dello stesso (o di un prodotto equivalente ed intercambiabile) per scongiurare eventuali problemi di approvvigionamento.

La situazione mutò improvvisamente con l'avvento del Pentium, Intel decise di mettere fine al fenomeno dei processori cloni e da allora i progettisti di CPU alternative ai prodotti Intel si sono dovuti arricciare le maniche per sviluppare da capo prodotti competitivi e al contempo totalmente compatibili.

### **AMD K5 - La risposta al Pentium**



Fu così che AMD propose il suo primo processore "indipendente": il K5. Il K5 era, sugli interi, migliore a parità di clock rispetto sia al Pentium che al Cyrix 6x86 ma fu introdotto in ritardo sul mercato esibendo minori frequenze rispetto alla concorrenza e prestazioni inferiori sul versante Floating-Point. Questo ritardo nell'introduzione di valide alternative al Pentium fece perdere molte quote di mercato ai concorrenti che dovettero negli anni successivi faticare molto per recuperare il terreno perduto (già nel '98 Intel deteneva l'87% del mercato!). A poco valse l'introduzione del P-Rating...Il K5 aveva 16K di I-cache L1 e 8K di D-cache L1 annoverando 4.3 milioni di transistor, massima freq. raggiunta: 166MHz

Modello	Data di rilascio	Tecnologia	Effettiva velocità
PR 75	27 Marzo 1996	0,5 micron	75 MHz
PR 90	27 Marzo 1996	0,5 micron	90 MHz
PR 100	7 Ottobre 1996	0,5 micron	100 MHz
PR 120	7 Ottobre 1996	0,35 micron	90 MHz
PR 133	7 Ottobre 1996	0,35 micron	100 MHz
PR 166	13 Gennaio 1997	0,35 micron	116.7 MHz

Nel frattempo una piccola società di nome NextGen aveva progettato un core compatibile x86 capace di decodificare le complesse istruzioni CISC dell'ISA x86 (vedi qui per maggiori chiarimenti sull'ISA) in istruzioni RISC più semplici. AMD pensò bene di acquisire la società e produrre in fretta e furia il K6, contemporaneamente Intel era già entrata nella fase Pentium Pro, Pentium II.

### AMD K6 - Il primo processore AMD di 6° generazione



Il K6 implementò l'architettura RISC86 superscalare, aggiunse il supporto alle istruzioni MMX, eliminò il P-Rating e portò la cache di primo livello a ben 32K + 32K (contro i 16 + 16 del Pentium MMX, Pentium Pro e Pentium II). L'architettura interna ricalca le features già descritte per i processori di sesta generazione (vedi anche schema più in basso):

- Advanced RISC86 superscalar microarchitecture
- Seven parallel execution units
- Multiple sophisticated x86 to RISC86 decoders
- Two level branch prediction
- Speculative execution Out-of-Order Execution
- 64K on-chip level one cache
- 32K instruction cache
- 32K writeback data cache
- MMX capability
- Socket 7 compatible
- 0.35 micron architecture

Benchè fosse dotato di features avanzate, il K6 aveva una pecca: era costruito su pipeline a bassa latenza a 6 stadi e 8,8 milioni di transistor, ottima per ridurre gli stalli ma difficile da far salire in frequenza, almeno rispetto ai 10 stadi del Pentium II. La floating point unit non è completamente pipelined e quindi esibisce performance ampiamente inferiori al Pentium. La cache di secondo livello è sempre saldata su piastra e funziona a 66Mhz contro la cache integrata su schedina dedicata del Pentium II e cloccata a metà frequenza del processore; questo rappresentò una debolezza ma anche un punto di forza della piattaforma K6 perchè permetteva il riutilizzo di piastre Socket7 (quelle del Pentium) ed in ogni caso costi minori rispetto alla proposta Intel.

Modello	Data di rilascio	Tecnologia [micron]
166,200 e 233 MHz	2 Aprile 1997	0.35
233 e 266 MHz	6 Gennaio 1998	0,25
300 MHz	7 Aprile 1998	0,25

### AMD K6-II - Il secondo processore AMD di 6° generazione

Nel giugno del 1998 AMD passa al K6-2 (9,3 milioni di transistor) che inaugura il Super Socket 7 a 100MHz e introduce la tecnologia 3DNow!. Si tratta di 21 nuove istruzioni multimediali che anticiperanno le successive SSE di Intel. La tecnologia 3DNow! introduce l'approccio SIMD (Single Instruction Multiple Data) anche con i numeri in virgola mobile (MMX opera solo sugli interi) e permette l'esecuzione di fino a 4 istruzioni su numeri Floating point a singola precisione (32bit). AMD pensava di compensare le scarse prestazioni della sua Floating point unit con la tecnologia 3DNow!, ma lo scarso supporto da parte degli sviluppatori fece presto riemergere le gravi debolezze in quel settore. Il K6-II ottenne comunque un discreto successo nella fascia entry-level del mercato ma al salire della frequenza cominciò a pesare la scarsa velocità della L2 cache esterna (100MHz).

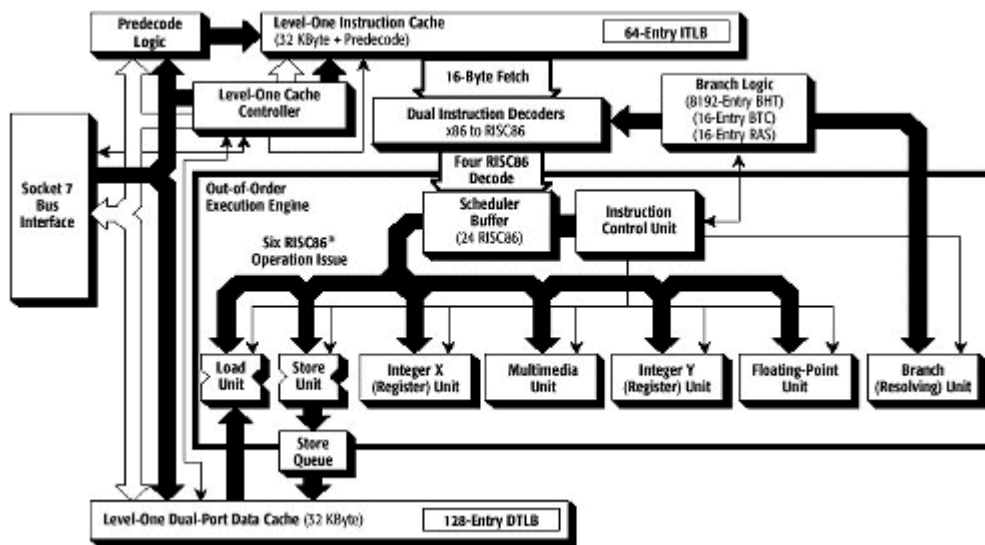
Modello	Data di rilascio
266, 300 e 333 MHz	28 Maggio 1998
350 MHz	27 Agosto 1998
366, 380 e 400 MHz	16 Novembre 1999
450 MHz	26 Febbraio 1999
475 MHz	5 Aprile 1999
500 MHz	30 Agosto 1999
533 MHz	29 Novembre 1999
550 MHz	22 Febbraio 2000

### AMD K6-III - L'ultimo esponente della famiglia K6



Nel '99 AMD introdusse il K6-3 che per un pò fece dimenticare i problemi del Super Socket 7 grazie ad una L2 cache da 256KByte integrata direttamente sul die del processore ed operante alla stessa frequenza. Questo stratagemma fece vedere di cosa era realmente capace il core K6-II e permise ad AMD di ottenere prestazioni sugli interi migliori del corrispondente P-III. Il 22 Febbraio del 1999 furono lanciati i modelli a 450 e 500MHz. Nel K6-3 AMD riuscì a integrare 21.300.000 transistor usando una tecnologia 0.25 micron. Ancora scarse le prestazioni sul versante Floating-Point.

Ecco infine lo schema funzionale del core K6, K6-II e K6-III:



- **Processori di settima generazione: AMD K7**



Intenzionata a superare tutti i limiti dei precedenti progetti, AMD produce un nuovo processore capace finalmente di competere su tutti i fronti con i prodotti Intel. Ed infatti con l'immissione sul mercato del K7 Athlon, per la prima volta, Intel viene battuta su tutti i fronti: sul fronte della massima frequenza di clock, sul fronte delle prestazioni velocistiche assolute e relative, sul fronte dei prezzi. E' un vero smacco per Intel. All'uscita del K7 Athlon a 600MHz Intel proponeva il P-III a "solo" 550MHz, ed inoltre il K7 era più veloce sia su gli interi che sul Floating Point rispetto al PIII grazie ad un redesign complessivo del core ma in particolar modo della unità Floating Point Unit, che fece salire il numero dei transistor a ben 22 Milioni (esclusa la L2 cache esterna).

Queste in breve le features salienti del K7:

- Decoders multipli per le istruzioni x86
- ICU a 72 ingressi
- Dynamic Branch prediction avanzata
- 3 unità di esecuzione floating point fuori sequenza totalmente pipelined (15 stadi) che eseguono tutte le istruzioni floating point x87 e istruzioni Mmx e 3dnow
- 3 unità superscalari per gli interi pipelined (10stadi) e con esecuzione fuori sequenza
- Tecnologia Enhanced 3dnow : nuove istruzioni per riconoscimento vocale, codifica video e scambio di dati per plug-ins internet e altre applicazioni di streaming video.
- Architettura della cache ad alte prestazioni: cache L1 ad alte prestazioni da 64KB + 64KB ed una interfaccia per la cache di 2° livello programmabile ad alta velocità.
- Bus di sistema a 200 Mhz derivato dall' EV6 di Alpha.

## L'architettura dell'Athlon

L' Athlon include 3 decoders per istruzione x86. Questi decoders traducono le istruzioni x86 in macro operations (MacroOPs) a lunghezza fissa per un più alto rendimento nell'esecuzione dell'elaborazione. Invece di eseguire direttamente le istruzioni x86 che hanno lunghezza da 1 a 15 bytes, l'Athlon esegue le MacroOPs RISC-Like migliorando di molto le prestazioni delle altre unità di elaborazione ed ottimizzazione.

Una volta che le MacroOPs sono decodificate, fino a 3 MacroOPs sono inviate all' ICU, per ogni ciclo di clock. L'ICU è un Buffer Reorder per MacroOPs a 72 entry che gestisce lo smistamento delle istruzioni, esegue la rinominazione del registro per gli operandi, e gestisce tutti gli stati d'eccezione e le operazioni di ritiro. L' ICU invia le MacroOPs agli Schedulers delle numerose unità di esecuzione multiple presenti nel K7.

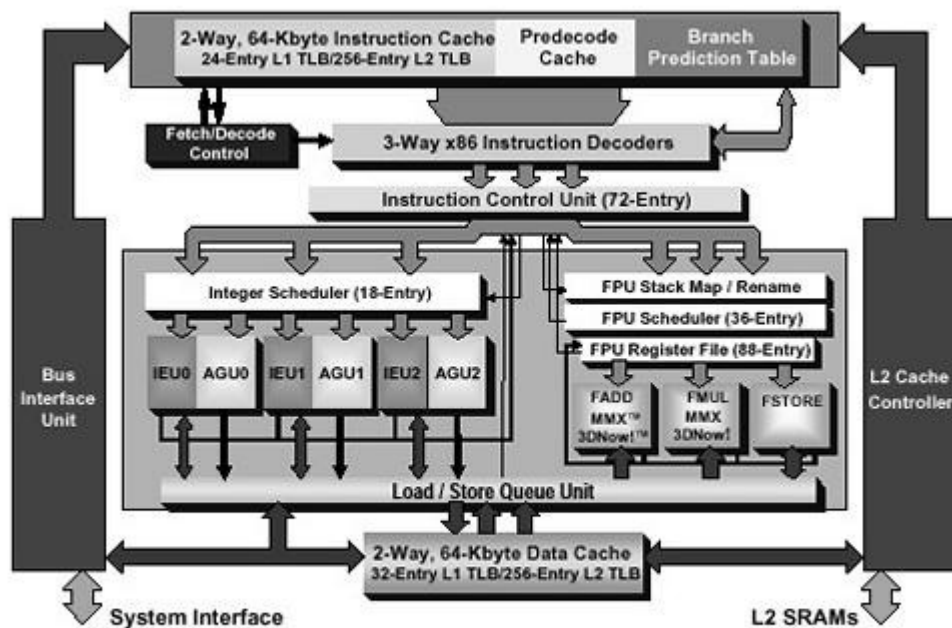
L' Athlon contiene uno scheduler a 18 entry per le istruzioni sui numeri interi e uno scheduler a 36 entry per l'FPU/3DNow. Questi schedulers distribuiscono le MacroOPs alle nove pipeline di esecuzione indipendenti

- 3 per i calcoli sugli interi
- 3 per il calcolo degli indirizzi
- 3 per l'esecuzione delle Mmx, 3dnow! e istruzioni floating point x87

L' Athlon offre il più potente e avanzato motore di floating point per piattaforma x86. L'FPU dell' Athlon è basata su 3 unità di esecuzione completamente pipelined (contro le due unità del PIII). Queste 3 unità di esecuzione (FMUL, FAD e FSTORE) eseguono tutte le istruzioni x87, Mmx, Enhanced 3dnow.

I primi Athlon furono costruiti con tecnologia a 0.25 micron e interconnessioni in Alluminio ed erano posti in una cartuccia tipo Pentium II e inseriti in uno slot chiamato Slot A, simile concettualmente allo Slot I di Intel. Nella cartuccia era presente una L2 cache di 512KB funzionante tipicamente ad 1/2 o 1/3 della frequenza del core.

Qui di seguito trovate lo schema esplicativo del core di tutta la famiglia K7:



AMD Athlon™ Processor Block Diagram

Modello	Data di rilascio	PT	FL2
500, 550, 600 MHz	29 Aprile 1999	0,25	1/2
650 MHz	9 Agosto 1999	0,25	1/2
700 MHz	4 Ottobre 1999	0,25	1/2
550, 600, 650, 700 MHz	29 Novembre 1999	0,18	1/2
750 MHz	29 Novembre 1999	0,25	2/5
800 MHz	6 Gennaio 2000	0,18	2/5
850 MHz	14 Febbraio 2000	0,18	2/5
900,950 MHz e 1 GHz	6 Marzo 2000	0,18	1/3

## AMD K7 - Thunderbird e Palomino



Come Intel, anche AMD passò più tardi (seconda metà 2000) all'integrazione di una cache di secondo livello direttamente sul die del processore. Con il passaggio alla tecnologia 0.18 micron con interconnessioni in Rame (core Thunderbird), AMD è riuscita a portare un consistente aumento delle frequenze e una sensibile diminuzione del calore prodotto oltre alla su menzionata integrazione di una L2 cache, da 256KByte per il modello Athlon e da 64KByte per il Duron, operante a piena velocità. Il passaggio coincidente anche con l'abbandono dello SlotA in favore del più economico e pratico SocketA.

Il resto è storia dei nostri giorni: Il P-III ha perso la sfida con il Thunderbird sia sotto il profilo delle frequenze (il PIII si è dovuto fermare a 1.1GHz per limiti tecnologici) che sotto quello della potenza specifica e del costo. Certamente AMD ha saputo sfruttare bene il grave momento di stallo che Intel vissuto dai tempi dell'fallimentare PIII 1.13GHz fino a poco tempo fa, giocando molto sulla leva prezzo per abbracciare sempre più ampie fette di mercato ai danni dell'avversario. Alla stabilizzazione della piattaforma Pentium 4 da parte di Intel, AMD risponde con una ulteriore ottimizzazione della collaudata ed efficientissima architettura K7: l'Athlon XP da poco uscito nei negozi. Poche le innovazioni del core Palomino ma sufficienti a guadagnare un incremento di prestazioni dell'ordine del 10-20% rispetto alle soluzioni precedenti e a consolidare la quota di mercato di AMD che pare sia ritornata su un valore del 30%. Con L'Athlon XP AMD reintroduce anche il vecchip P-Rating come risposta alle elevate frequenze del Pentium4 non rappresentative della potenza sviluppata in relazione ai prodotti AMD.

Come preannunciato, AMD non ha introdotto grosse modifiche al suo nuovo core (in particolare le unità di elaborazione principali e le cache rimangono identiche a quelle precedenti) ma ne ha ottimizzato alcuni aspetti per poter ottenere minori consumi, maggiori velocità di clock e maggiore potenza utilizzando lo stesso numero di transistor (37,5 Milioni contro i precedenti 37 Milioni del Thunderbird). Andiamo ad analizzare queste nuove caratteristiche:

### Incremento delle L1 TLB Entries

Si tratta di una piccola cache interna al processore che viene utilizzata per accelerare il processo di traslazione degli indirizzi da logici a fisici ( Translation Lookaside Buffer ). L'incremento del buffer porta a una maggiore uniformità delle prestazioni in condizioni di Multi-Tasking e nei Server (modello Athlon MP).

## **Introduzione del Data Prefetch**

Si tratta di un meccanismo di previsione dei dati che verranno utilizzati dal flusso di istruzioni in esecuzione; la predizione permette di caricare in cache anticipatamente i dati necessari aumentando il rendimento nell'accesso alla memoria esterna. I maggiori benefici si hanno in architetture ad elevata banda e alta latenza quindi quelle tipiche del Pentium 4 (che infatti ha un suo meccanismo avanzato di pre-fetch) e dell'Athlon con memoria DDR. La vera novità è comunque rappresentata dalla possibilità di controllare via software il Prefetch.

## **Compatibilità con le SSE**

AMD ha introdotto nel set 3DNow 52 nuove istruzioni che rendono di fatto l'Athlon completamente compatibile con la tecnologia SSE di Intel. Ai puristi questo può sembrare un controsenso ma permetterà agli sviluppatori di uniformare lo sviluppo del codice sia per Athlon che per Pentium. L'implementazione del set di istruzioni SSE2 (presenti nel P4) sarà appannaggio della futura architettura x86-64 di AMD.

## **Tecnologia PowerNow!**

La tecnologia di riduzione del consumo permette all'Athlon un utilizzo proficuo nel settore Mobile (modello Athlon 4). PowerNow! consente di variare tensione e frequenza del core in funzione dell'impiego del processore. Sono previsti fino a 32 step a partire da 500MHz - 1.2V fino alla frequenza massima del processore. La tensione del core alla frequenza nominale si è molto ridotta passando dagli 1.7V a 1.4V (per frequenze intorno al GHz). AMD ha pensato bene di inserire anche un diodo di relavamento termico (analogamente al PIII e PIV) per garantire una maggiore protezione del core da eventuali stress termici. La riduzione dei consumi è al minimo del 20% (grazie alla tensione del core), mentre la tecnologia PowerNow! permette di estendere l'autonomia delle batterie di un portatile di fascia media fino a 3 - 3,5 ore.