UNIVERSITÀ DEGLI STUDI DI PADOVA
FACOLTÀ DI INGEGNERIA

Corso di Laurea in Ingegneria Elettronica

Tesi Elaborato scritto

EVOLUZIONE E STATO DELL'ARTE DEI MICROPROCESSORI INTEL
INTEL MICROPROCESSORS EVOLUTION AND STATE OF THE ART

RELATORE:
PROF. SERGIO CONGIU

LAUREANDO:
NICOLA FAVERO
Matr.: 153279

ANNO ACCADEMICO 2010-2011

Indice

Capitolo I˚. Microarchitettura di un microprocessore

§ 1 - Definizione di Microarchitettura

§ 2 - Aspetti principali delle Microarchitetture

§ 2.1 - Pipeline dati

§ 2.2 - Problematiche della pipeline

§ 2.3 - Evoluzione della pipeline

§ 2.4 - Scelta del set di istruzioni

§ 2.5 - Cache

§ 2.6 - BPU (Branch Prediction Unit)

§ 2.7 - Architettura superscalare

§ 2.8 - Esecuzione fuori ordine

§ 2.9 - Multiprocessing

§ 2.10 - Multithreading

Capitolo II˚. Evoluzione dei microprocessori Intel (Roadmap)

§ 1 - Il primo Microprocessore

§ 2 - La prima Microarchitettura

Capitolo III˚. Tecnologie di produzione

§ 1 - Evoluzione della tecnologia del processo di fabbricazione

§ 1.1 - Tecnologia 65nm

§ 1.2 - Tecnologia 45nm

§ 1.3 - Tecnologia 32nm

§ 1.4 - Tecnologia 22nm

Capitolo IV˚. Microarchitetture moderne

§ 1 - Microarchitetture recenti

§ 1.1 - Processori di ottava generazione (2006 - Core 65nm)

§ 1.1.1 - Evoluzione di Core (2008 - Penryn 45nm)

§ 1.2 - Processori di nona generazione (2008 - Nehalem 45nm)

§ 1.2.1 - Innovazioni della microarchitettura Nehalem
§ 1.2.2 - Evoluzione della Pipeline nella microarchitettura Nehalem
§ 1.2.3 - Evoluzione di Nehalem (2010 - Westmere 32nm)
§ 2 - Microarchitetture attuali
§ 2.1 - Processori di decima generazione (2010 – Sandy Bridge 32nm)
§ 3 - Microarchitetture future
§ 3.1 - Evoluzione di Sandy Bridge (2012 – Ivy Bridge)
§ 3.2 - Processori di undicesima generazione (2013 - Haswell 22nm)
§ 4 – Conclusioni

INTRODUZIONE

A partire dalla invenzione del primo transistor nel 1947 ad opera di Shockeley, Bardeen e Brattain la tecnologia elettronica ha progredito rapidamente aprendo la strada alla creazione di prodotti più avanzati e potenti e al tempo stesso meno costosi e più efficienti. Nonostante questi miglioramenti il calore dissipato dai transistor e le correnti di dispersione rimangono delle barriere critiche alla costruzione di transistor sempre più piccoli e al perpetuarsi della legge di Moore.

Nel 1971 con l'invenzione del primo microprocessore, il 4004, da parte di Faggin, Hoff e Mazor di Intel inizia un percorso tecnologico che a partire dai 2300 transistor che lo componevano e da un processo produttivo a 10 micron arriva all'attuale tecnologia di produzione a 32 nanometri (nm) con densità circuitale di quasi 1 miliardo di transistor per circuito integrato.

Questa tesi di laurea si pone l'obiettivo di descrivere l'evoluzione delle architetture dei processori Intel esaminando la Roadmap passata, presente e futura, considerando le tecnologie del processo di costruzione ed esaminando nello specifico le architetture recenti di 45nm, quella attuale a 32nm e l'ormai prossima a 22nm.

Verrà evidenziato come l'introduzione di nuovi materiali e di nuove tipologie di transistor abbia permesso l'eliminazione delle barriere che tendevano a limitare la progressione della legge di Moore portando alla costruzione di processori che, utilizzando sempre meno energia, consentono la creazione di prodotti dal laptop al pc desktop al server sempre meno costosi e sempre più performanti.

Capitolo 1. Microarchitettura di un microprocessore

§ 1 - Definizione di Microarchitettura

Con **Microarchitettura** si intende il metodo in cui un dato insieme di istruzioni (instruction set architecture **ISA**) è implementato all'interno di un processore. Un set di istruzioni può essere implementato in maniera diversa in differenti microarchitetture in base agli obiettivi dei costruttori o all'evolversi della tecnologia pur essendo in grado di eseguire gli stessi programmi. Nuove microarchitetture e/o soluzioni circuitali insieme all'evoluzione dei processi di costruzione dei semiconduttori consentono di ottenere performance sempre superiori con i nuovi microprocessori pur usando sempre lo stesso set di istruzioni (ISA).[1]

L'architettura di un computer comprende almeno tre sottocategorie :

- L'**Instruction Set Architecture (ISA)** come l'immagine astratta di un sistema di elaborazione così come è vista da un programmatore in assembly (linguaggio macchina) includendo il set di istruzioni, la lunghezza della parola, il modo di indirizzamento della memoria, i registri di processo, gli indirizzi e i formati dei dati. [2]
- La **Microarchitettura** è al livello più basso più concreto e più dettagliato il modo in cui le parti del microprocessore sono interconnesse tra di loro e come operano per implementare il set di istruzioni (ISA). Per essere più precisi, ad esempio, la dimensione della memoria cache del microprocessore è un aspetto organizzativo che generalmente non ha nulla a che fare con l'ISA.
- Il **System Design** include tutte le altre componenti di un sistema di elaborazione come bus di sistema, controller di memoria, meccanismi di scarico del lavoro della CPU come la direct memory access (DMA) e problematiche come il multiprocessing.

[1] http://en.wikipedia.org/wiki/Microarchitecture
[2] http://en.wikipedia.org/wiki/Computer_architecture

Una volta specificate sia l'ISA che la microarchitettura, si passa alla progettazione hardware della CPU a livello logico,circuitale e fisico.

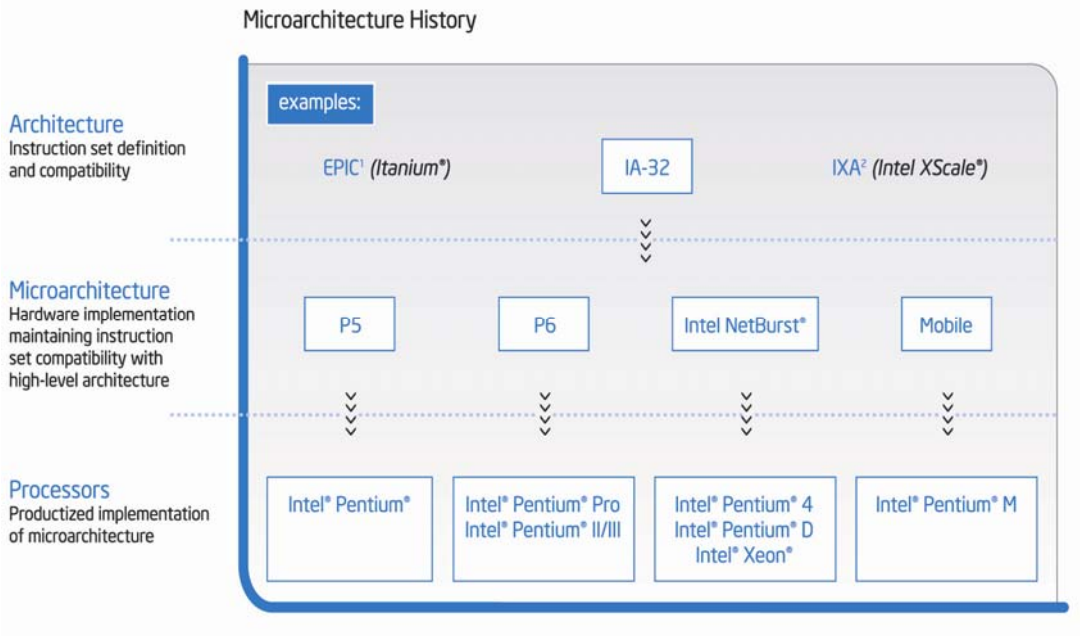

Figura 1 Relazione tra architettura del processore e microarchitettura.

Con Architettura del Processore si intende l'insieme del set di istruzioni, dei registri, e delle strutture di dati residenti in memoria che sono visibili ad un programmatore. L'architettura del processore mantiene la compatibilità con il set di istruzioni in modo che il processore possa elaborare codice scritto per le generazioni di processori passate, presenti e future. Con Microarchitettura si intende invece l'implementazione dell'architettura del processore nel silicio. All'interno della stessa famiglia di processori la microarchitettura è spesso soggetta ad evoluzione per aumentare le prestazioni e la capacità mantenendo la compatibilità con l'architettura.

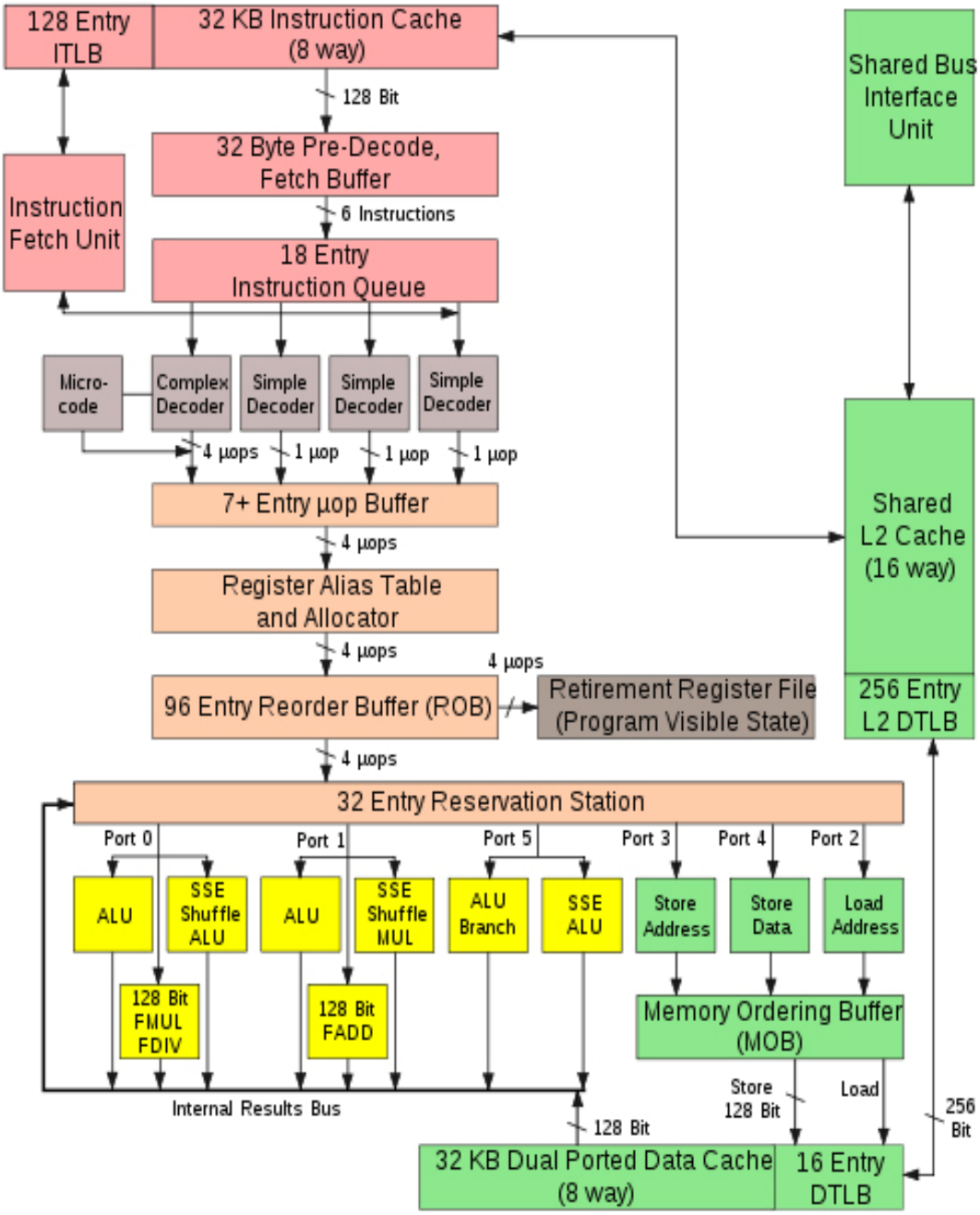

Figura 2 Esempio di Microarchitettura di un microprocessore Intel Core 2 (2006)

§ 2 – Aspetti principali delle Microarchitetture

§ 2.1 – Pipeline dati

L'implementazione della **Pipeline** è uno degli aspetti principali di ogni microarchitettura. La **pipeline dati** è una tecnologia utilizzata nell'architettura hardware dai microprocessori dei computer per incrementare il throughput, ovvero la quantità di istruzioni eseguite in una data quantità di tempo, parallelizzando i flussi di elaborazione di più istruzioni. [3]

L'elaborazione di un'istruzione da parte di un processore si compone di cinque passaggi fondamentali:

1. **IF** (Instruction Fetch): Lettura dell'istruzione da memoria
2. **ID** (Instruction Decode): Decodifica istruzione e lettura operandi da registri
3. **EX** (Execution): Esecuzione dell'istruzione
4. **MEM** (Memory): Attivazione della memoria (solo per certe istruzioni)
5. **WB** (Write Back): Scrittura del risultato nel registro opportuno

Praticamente ogni CPU in commercio è gestita da un clock centrale e ogni operazione elementare richiede almeno un ciclo di clock per poter essere eseguita. Le prime CPU erano formate da un'unità polifunzionale che svolgeva in rigida sequenza tutti e cinque i passaggi legati all'elaborazione delle istruzioni. Una CPU classica richiedeva quindi almeno cinque cicli di clock per eseguire una singola istruzione.

Figura 1 Cpu senza pipeline

Con il progresso della tecnologia si è potuto integrare un numero maggiore di transistor in un microprocessore e quindi si sono potute parallelizzare alcune

[3] http://it.wikipedia.org/wiki/Pipeline_dati

operazioni riducendo i tempi di esecuzione. La pipeline dati rappresenta la massima parallelizzazione del lavoro di un microprocessore.

Una CPU con pipeline è composta da cinque stadi specializzati, capaci di eseguire ciascuno una operazione elementare di quelle sopra descritte. La CPU lavora come in una catena di montaggio cioè ad ogni stadio provvede a svolgere in maniera sequenziale un solo compito specifico per l'elaborazione di una certa istruzione. Quando la catena è a regime, ad ogni ciclo di clock dall'ultimo stadio esce un'istruzione completata. Nello stesso istante ogni unità sta però elaborando in parallelo i diversi stadi di successive altre istruzioni. In sostanza quindi si guadagna una maggior velocità di esecuzione a prezzo di una maggior complessità circuitale del microprocessore, che non deve essere più composto da una sola unità, ma da cinque unità che devono collaborare tra loro.

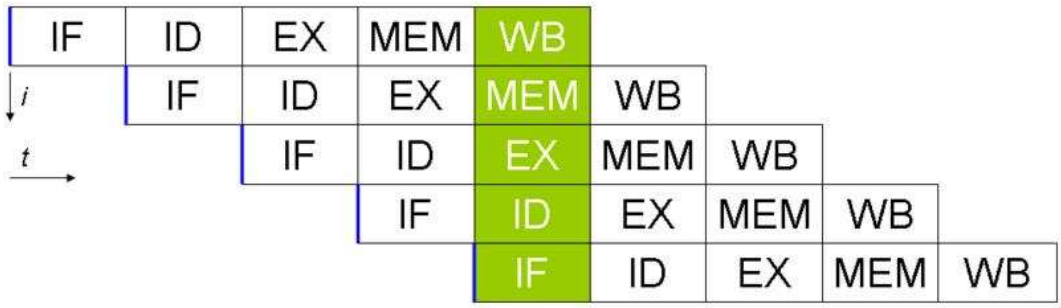

Figura 2 Cpu con pipeline a 5 stadi

§ 2.2 – Problematiche della pipeline

L'implementazione di una pipeline non sempre moltiplica il throughput finale. L'analisi delle problematiche legate alla gestione delle pipeline per ottenere le migliori prestazioni teoriche ricadono sotto la ricerca del parallelismo a livello d'istruzione (instruction level parallelism), cioè le istruzioni che possono essere eseguite in parallelo senza creare conflitti o errori di esecuzione. Principalmente le singole pipeline affrontano due problemi; il problema legato alla presenza di istruzioni che possono richiedere l'elaborazione di dati non ancora disponibili e il problema legato alla presenza di salti condizionati.

- Il primo problema deriva dal lavoro parallelo delle unità.

Supponiamo che la CPU con pipeline debba eseguire il seguente frammento di codice:

1. C=A+B
2. D=C-1

La prima istruzione deve prelevare i numeri contenuti nelle variabili A e B, sommarli e porli nella variabile C. La seconda istruzione deve prelevare il valore contenuto nella variabile C, sottrarlo di uno e salvare il risultato in D. Ma la seconda istruzione non potrà essere elaborata (EX) fino a quando il dato della prima operazione non sarà disponibile in memoria (MEM) e quindi la seconda operazione dovrà bloccarsi per attendere il completamento della prima e quindi questo ridurrà il throughput complessivo.

- Il secondo problema consiste nei salti condizionati.

I programmi contengono delle istruzioni condizionate che se una specifica condizione logica è verificata provvedono a interrompere il flusso sequenziale del programma e a mandare in esecuzione un altro pezzo di programma indicato dall'istruzione di salto. Ogni volta che questo accade il microprocessore si trova a dover eseguire un nuovo flusso di operazioni e quindi deve svuotare la pipeline del precedente flusso e caricare il nuovo flusso. Ovviamente queste operazioni fanno sprecare cicli di clock e quindi deprimono il throughput. Per ridurre questo problema le CPU adottano delle unità chiamate unità di predizione delle diramazioni (in inglese *Branch Prediction Unit*) che fanno delle previsioni sul flusso del programma. Queste unità riducono notevolmente i cicli persi per i salti.

§ 2.3 – Evoluzione della pipeline

Per realizzare CPU con prestazioni migliori col tempo si è affermata la strategia di integrare in un unico microprocessore più pipeline che funzionano in parallelo. Questi microprocessori sono definiti superscalari dato che sono in grado di eseguire mediamente più di un'operazione per ciclo di clock. Queste pipeline ovviamente rendono ancora più complessa la gestione dei problemi di coerenza e dei salti condizionati. Nelle CPU moderne inoltre le pipeline non sono composte da soli cinque stadi ma in realtà ne utilizzano molti di più. Questo si è reso necessario per potere innalzare la frequenza di clock. Spezzettando le singole operazioni necessarie per completare un'istruzione in tante sotto operazioni si può elevare la frequenza della CPU dato che ogni unità deve svolgere un'operazione più semplice e quindi può impiegare meno tempo per completare la sua operazione. Questa scelta di progettazione consente effettivamente di aumentare la frequenza di funzionamento delle CPU, ma rende critico il problema dei salti condizionati. In caso di un salto condizionato non previsto un processore con 20 stadi di pipeline per esempio può essere costretto a svuotare e ricaricare una pipeline di 20 stadi perdendo fino a 20 cicli di clock contro una classica CPU a pipeline a 5 stadi che avrebbe sprecato nella peggiore delle ipotesi 5 cicli di clock.

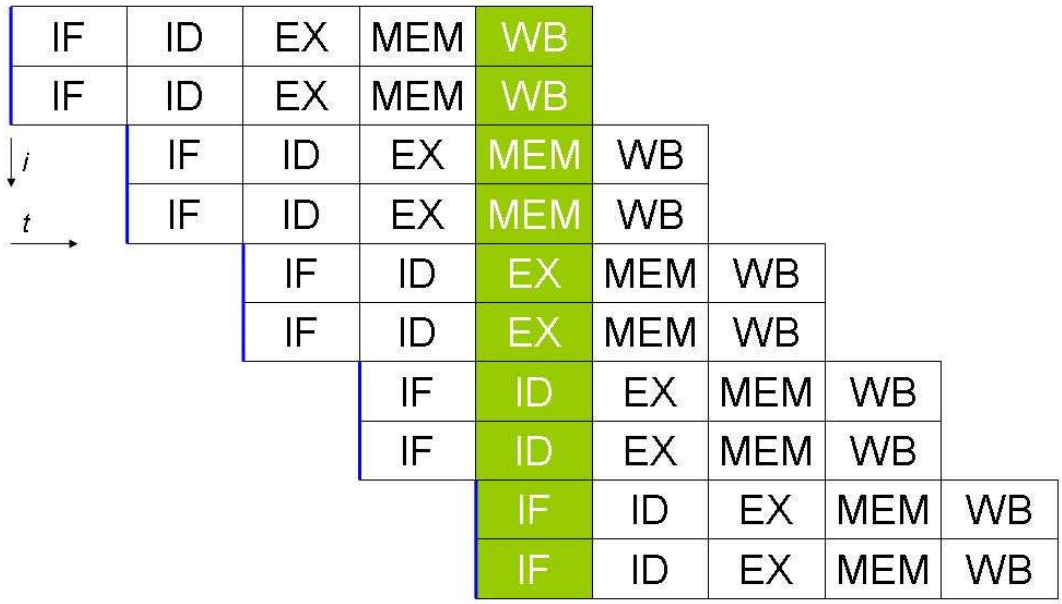

Figura 3 Cpu superscalare a doppia pipeline

La sempre maggior richiesta di potenza di calcolo ha spinto le industrie produttrici di microprocessori a integrare in un unico chip più microprocessori (architetture a multiprocessore). Questo strategia consente al computer di avere due CPU separate dal punto di vista logico, ma fisicamente risiedenti nello stesso chip. Questa strategia progettuale attenua i problemi di coerenza e di predizione dei salti. Infatti ogni CPU logica esegue un programma separato e quindi tra i diversi programmi non si possono avere problemi di coerenza tra le istruzioni. Questa scelta progettuale aumenta le prestazioni solo nel caso in cui il sistema operativo sia in grado di utilizzare più programmi contemporaneamente e i programmi siano scritti per poter utilizzare tutte le CPU disponibili, cioè se i programmi sono parallelizzabili.

§ 2.4 – Scelta del set di istruzioni

Altro fattore determinante per l'ottimizzazione delle prestazioni di un microprocessore è la scelta del tipo del set di istruzioni implementate.
Quando i transistor disponibili su un solo chip erano pochi e i calcolatori venivano spesso programmati in assembly, era naturale sfruttarli in modo tale da avere CPU con istruzioni potenti, evolute e complesse: più queste erano vicine alle istruzioni dei linguaggi di programmazione ad alto livello più il computer sarebbe stato facile da programmare, e i programmi avrebbero occupato poco spazio in memoria (anch'essa poca e preziosa). Le CPU progettate secondo questo approccio sono dette **CISC** (Complex Instruction Set Computer) ed avevano unità di controllo complesse capaci di sfruttare al meglio pochi registri e i cui programmi erano di dimensioni relativamente piccole. A cavallo fra gli anni '70 e gli '80 la situazione però cambiò rapidamente: la RAM divenne più economica e comparvero i primi compilatori moderni, ottimizzanti, in grado di generare linguaggio macchina molto efficiente: per questo si iniziò a pensare ad un nuovo modo di progettare le CPU, prendendo in esame la possibilità di usare i transistor disponibili per avere invece molti registri e un set di istruzioni elementare, molto ridotto, che delegasse al compilatore il lavoro di tradurre le istruzioni complesse in serie di istruzioni più semplici, permettendo così di avere unità di controllo particolarmente semplici e veloci. Le CPU progettate secondo questo approccio sono dette **RISC** (Reduced Instruction Set Computer). Attualmente la distinzione fra queste due classi di architetture è venuta in gran parte

meno: il numero di transistor disponibili su un solo chip è aumentato tanto da poter gestire molti registri ed anche set di istruzioni complesse.[4]

§ 2.5 – Cache

La **CPU cache** è la cache utilizzata dalla CPU di un computer per ridurre il tempo medio d'accesso alla memoria. La cache è un tipo di memoria piccola, ma molto veloce, che mantiene copie dei dati ai quali si fa più frequentemente accesso in memoria principale. Finché la maggior parte degli accessi alla memoria avviene su dati caricati nella cache, la latenza media dell'accesso alla memoria sarà più vicina alla latenza della cache piuttosto che a quella della memoria principale.

Quando il processore vuole leggere o scrivere in una data collocazione in memoria principale, inizialmente controlla se il contenuto di questa posizione è caricato in cache. Questa operazione viene effettuata confrontando l'indirizzo della posizione di memoria con tutte le etichette nella cache che potrebbero contenere quell'indirizzo. Se il processore trova che la posizione di memoria è in cache, si parla di *cache hit* (accesso avvenuto con successo), altrimenti di *cache miss* (fallimento d'accesso). Nel caso di un *cache hit*, il processore legge o scrive immediatamente il dato sulla linea di cache. Il rapporto tra *cache hit* e accessi totali è chiamato anche *hit rate* ed è una misura dell'efficacia della cache stessa.

Nel caso di un *cache miss*, la maggior parte delle cache crea una nuova entità, che comprende l'etichetta appena richiesta dal processore ed una copia del dato dalla memoria. Un fallimento del genere è relativamente lento, in quanto richiede il trasferimento del dato dalla memoria principale, il cui tempo di risposta è molto maggiore di quello della memoria cache.[5]

Il grande vantaggio dei vari tipi di cache presenti nella cpu è legato all'uso di questo tipo di memoria con il meccanismo di pipeline. All'inizio non aveva molto senso utilizzare una pipeline che poteva elaborare istruzioni molto più velocemente del

[4] http://it.wikipedia.org/wiki/CPU
[5] http://it.wikipedia.org/wiki/CPU_cache

tempo di latenza della memoria principale. Utilizzando invece la cache la pipeline può elaborare le istruzioni con il tempo di latenza della cache che è molto più breve del tempo di latenza della memoria ram. Questo fatto ha permesso di innalzare le frequenze di lavoro dei processori molto più di quelle delle memorie di sistema.

§ 2.6 – BPU (Branch Prediction Unit)

La **predizione delle diramazioni** (*branch prediction*) è il compito della **BPU** (*Branch Prediction Unit*), una componente della CPU che cerca di prevedere l'esito di un'operazione su cui si basa l'accettazione di una istruzione di salto condizionato, evitando rallentamenti che possono essere molto evidenti in una architettura con pipeline.

L'importanza di questa operazione è evidente soprattutto per i microprocessori moderni, superscalari e con lunghe pipeline, che per ogni errore di previsione devono sprecare molti cicli di clock di lavoro prezioso.

La predizione delle diramazioni non va confusa con la predizione dell'indirizzo di arrivo della diramazione. Questa predizione viene svolta dall'unità branch target predictor che cerca di predire l'indirizzo di arrivo del salto e di caricare le istruzioni corrispondenti prima che il salto sia svolto in modo da evitare rallentamenti dovuti al caricamento delle istruzioni dopo il salto.[6]

§ 2.7 – Architettura superscalare

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 tutte le CPU *general purpose* prodotte dal 1998.[7]

I microprocessori più semplici sono scalari: ogni operazione eseguita tipicamente manipola uno o due operandi contemporaneamente. Invece, in un processore

[6] http://it.wikipedia.org/wiki/Predizione_delle_diramazioni
[7] http://it.wikipedia.org/wiki/Microprocessore_superscalare

vettoriale, una singola istruzione viene applicata su di un vettore, formato da più dati raggruppati. In questo modo, una applicazione che deve eseguire una certa operazione su una grande quantità di dati viene svolta con molta più rapidità. Un processore superscalare è una forma intermedia tra i due: istruzioni diverse trattano i propri operandi contemporaneamente, su diverse unità hardware all'interno dello stesso chip. In questo modo più istruzioni possono essere eseguite nello stesso ciclo di clock.

In una CPU superscalare sono presenti diverse unità funzionali dello stesso tipo, con dispositivi addizionali per distribuire le istruzioni alle varie unità. Per esempio, sono generalmente presenti numerose unità per il calcolo intero (definite ALU). Le unità di controllo stabiliscono quali istruzioni possono essere eseguite in parallelo e le inviano alle rispettive unità. Questo compito non è facile, dato che un'istruzione può richiedere il risultato della precedente come proprio operando, oppure può dover impiegare il dato conservato in un registro usato anche dall'altra istruzione; il risultato può quindi cambiare secondo l'ordine d'esecuzione delle istruzioni. La maggior parte delle CPU moderne dedica molta potenza per svolgere questo compito con la massima precisione possibile, per permettere al processore di funzionare a pieno regime in modo costante; compito che si è reso sempre più importante con l'aumento del numero delle unità. Mentre le prime CPU superscalari possedevano due ALU ed una FPU un processore attuale può avere una molteplicità di ALU (unità logico aritmetiche), più FPU (unità floating point) e varie unità SIMD (Single instruction, multiple data**)**. Se il sistema di distribuzione delle istruzioni non mantiene occupate tutte le unità funzionali del processore, le sue prestazioni ne soffrono grandemente.

Attualmente è impensabile un futuro miglioramento sensibile del sistema di controllo, ponendo di fatto un limite ai miglioramenti prestazionali dei processori superscalari. Il progetto VLIW (very long instruction word) cerca una soluzione scaricando parte del processo di controllo delle istruzioni in fase di scrittura del programma e di compilazione, evitando al processore di doverlo ripetere ad ogni esecuzione del programma.

§ 2.8 – Esecuzione fuori ordine

L'**esecuzione fuori ordine** (out of order execution) indica la capacità di molti processori di eseguire le singole istruzioni senza rispettare necessariamente l'ordine imposto dal programmatore. Il processore in sostanza analizza il codice che dovrà eseguire, individua le istruzioni che non sono vincolate da altre istruzioni e le esegue in parallelo. Questa strategia permette di migliorare le prestazioni dei moderni microprocessori dal momento che si possono verificare casi in cui il processore rimane in stallo in attesa della risoluzione dei cache miss.
Nei primi processori un cache miss forzava il cache controller a porre il processore in stallo per aspettare che i dati necessari fossero prelevati dalla memoria. Dal momento che nella cache ci sono anche altre istruzioni di programma che possono essere utilizzate, l'esecuzione fuori ordine esamina le istruzioni presenti nella cache e le esegue comunque se non ci sono dipendenze con l'istruzione responsabile del cache miss e successivamente riordina i risultati ripristinando la sequenza del programma.

§ 2.9 – Multiprocessing

Con multiprocessing si definisce l'utilizzo di due o più CPU all'interno di un computer.[8]
Il termine si riferisce anche alla capacità del sistema di supportare due o più processori e/o la capacità di allocare compiti tra essi. Ci sono molte variazioni di questo aspetto e la definizione di multiprocessing può variare con il contesto principalmente in funzione di come sono definite le CPU. Si parla allora di core multipli sullo stesso die (inteso come la piastra di silicio su cui è realizzata la cpu), di die multipli nello stesso package, di package multipli nella stessa unità di sistema etc.
In un sistema multiprocessore tutte le cpu possono essere uguali o alcune possono essere riservate per compiti particolari. La progettazione dell'hardware e del software determinano la simmetria o meno del sistema multiprocessore. Per esempio possiamo impiegare sempre la stesso processore per rispondere agli interrupt hardware oppure

[8] http://en.wikipedia.org/wiki/Multiprocessing

utilizzare un unico processore per l'esecuzione di tutto il codice in kernel mode laddove il codice in user mode può essere eseguito da tutti gli altri processori.
Sistemi dove tutti i processori sono utilizzati allo stesso livello sono detti **SMP** (symmetric multiprocessing) .
Sistemi dove le cpu non sono utilizzate allo stesso modo e le risorse di sistema sono divise tra le varie cpu si identificano in **ASM**P (asymmetric multi processing), **NUM**A (non-uniform memory access) e clustered multi processing.

§ 2.10 – Multithreading

Il **multithreading** indica il supporto hardware da parte di un processore di eseguire più thread (sequenze di istruzioni)[9]. Questa tecnica si distingue da quella alla base dei sistemi multiprocessore per il fatto che i singoli thread condividono lo stesso spazio d'indirizzamento, la stessa cache e lo stesso TLB (translation lookaside buffer) (buffer utilizzato dalla MMU per velocizzare la traduzione degli indirizzi virtuali). Il multithreading migliora le prestazioni dei programmi solamente quando questi sono stati sviluppati suddividendo il carico di lavoro su più thread che possono essere eseguiti in parallelo. I sistemi multiprocessore sono dotati di più unità di calcolo indipendenti, un sistema multithread invece è dotato di una singola unità di calcolo che si cerca di utilizzare al meglio eseguendo più thread nella stessa unità di calcolo. Le tecniche sono complementari, a volte i sistemi multiprocessore implementano anche il multithreading per migliorare le prestazioni complessive del sistema.
La necessità del multithreading si è evidenziata quando ci si è posti il problema di utilizzare con la massima efficienza possibile le unità di calcolo. Si è appurato che molti programmi erano composti da più thread paralleli o potevano essere scomposti in più thread paralleli con lievi modifiche al codice sorgente. Quindi migliorando l'esecuzione di thread paralleli si poteva migliorare l'esecuzione complessiva dei programmi. Questo ha spinto lo sviluppo dei sistemi multithreading e dei sistemi multiprocessore.

[9] http://it.wikipedia.org/wiki/Multithreading

Le principali critiche al multithreading sono:

- Più thread condividono le stesse risorse come cache o translation lookaside buffer e quindi possono interferire a vicenda rallentandosi.
- Le prestazioni dei singoli thread non migliorano, ma anzi possono degradare all'aumento dei thread concorrenti.
- Il supporto hardware del multithreading e dei sistemi multiprocessori richiede anche un contributo del software, i programmi e i sistemi operativi devono essere adattati per gestire questa nuova possibilità.

Capitolo 2. Evoluzione dei microprocessori Intel (Roadmap)

§ 1 – Il primo Microprocessore

1971

Viene commercializzato il primo microprocessore, il **4004**, contenente poco più di **2300 transistor** costruito con tecnologia PMOS a **10 micron (10000 nm)**.

Specifiche tecniche

- Massima frequenza di clock di **0.740 MHz**
- Memorizzazione separata di codice e dati, il 4004 utilizza un singolo bus a 4 bit multiplexato per trasferire:
    - Indirizzi a 12 bit
    - Istruzioni in word di 8 bit, in uno spazio separato rispetto ai dati
    - Dati in word di 4 bit
- Il set di istruzioni comprende 46 istruzioni (di cui 41 a 8 bit e 5 a 16 bit)
- 16 registri a 4 bit
- Alimentazione a 12 Volt.
- Stack per le subroutine con al massimo 3 livelli di annidamento
- Poteva indirizzare fino a 640 Byte di memoria RAM.

A differenza dei microprocessori contemporanei, il 4004 includeva anche il controllo dei bus di memoria e di I/O che non sono normalmente gestiti dal microprocessore. Pertanto il 4004 non solo era una CPU completa, ma aveva anche funzionalità addizionali che normalmente non sono considerate compito della CPU.[10]

[10] http://it.wikipedia.org/wiki/Intel_4004

1972

Viene creato l'**8008**, una CPU a 8 bit con un bus indirizzi a 14 bit capace di indirizzare fino a 16 KB di memoria. Contiene **3500** transistor ed è costruito con tecnologia PMOS a **10 micron**.

Con una frequenza di clock di **0.8 Mhz** l'8008 risultava più lento del 4004 in termini di istruzioni per secondo (36000 contro 80000) ma il fatto che l'8008 potessse elaborare dati 8 bit alla volta gli permetteva di accedere a molta più RAM rendendolo dalle 3 alle 4 volte più veloce dei processori a 4 bit.[11]

1974

L'Intel **8080** è il successore dell'Intel 8008 (con cui è compatibile a livello di codice assembly). Contiene **6000** transistor, ha clock a **2 Mhz** ed è costruito con tecnologia nMOS a **6 micron**.

Il suo packaging DIP a 40 pin permette all'8080 di fornire un bus di indirizzi a 16 bit e un bus di dati a 8 bit, che consentono di accedere facilmente a 64 kilobyte di memoria. All'interno è dotato di sette registri a 8 bit (sei dei quali possono essere combinati a formare tre registri da 16 bit), uno stack pointer a 16 bit (che, al contrario di quanto accade nell'8008 che fa uso di una stack interna, punta in memoria), e un program counter a 16-bit.

L'8080 dispone di 256 porte I/O che permettono alle periferiche un collegamento senza allocazioni di memoria (come avviene per le periferiche mappate in memoria); uno svantaggio di questo sistema è la necessità di introdurre istruzioni aggiuntive per l'I/O. Il primo microcomputer a scheda singola fu costruito sulla base dell'8080.[12]

L'8080 è stato usato in molti computer storici, come Altair 8800 della MITSe l'IMSAI 8080, che tra i primi hanno eseguito il sistema operativo CP/M; un sistema che ha fruttato molto al successivo processore Zilog Z80, completamente compatibile con l'8080 e più potente: l'accoppiata Z80 - CP/M divenne infatti la combinazione

[11] http://en.wikipedia.org/wiki/Intel_8008
[12] http://en.wikipedia.org/wiki/Intel_8080

CPU/OS dominante, in modo simile con ciò che è accaduto un decennio dopo tra x86 ed MS-DOS.[13]

§ 2 – La prima Microarchitettura

1978 – Processori di prima generazione

L' **8086** è il primo microprocessore a 16 bit commercializzato e viene identificato come capostipite dell'**architettura x86** che vede susseguirsi varie generazioni fino all'attuale. Contiene **fino a 29000** transistor a seconda delle versioni, ha un clock a da **4,77 a 10 Mhz** ed è costruito con tecnologia nMOS a **3.2 micron** chiamata HMOS (High performance MOS).[14]

È basato sull'8080 (è compatibile con l'assembly dell'8080), con un insieme di registri simili, ma a 16 bit. L'unità di interfaccia con il bus (detta **BIU** da **B**us **I**nterface **U**nit) passa le istruzioni all'unità di esecuzione (detta **EU** da **E**xecution **U**nit) attraverso una coda FIFO di prefetch (6 byte), in modo che il fetch e l'esecuzione delle istruzioni sia contemporaneo – una forma primitiva di pipelining (le istruzioni dell'8086 hanno una dimensione tra 1 e 4 byte).

Ha quattro registri a 16 bit per uso generico, a cui si può accedere come se fossero otto registri a 8 bit, e quattro registri a 16 bit di indice (incluso lo stack pointer). Ha uno spazio di indirizzamento a 16 bit per l'I/O (cioè può accedere a 65.536 dispositivi di I/O a 8 bit) e dispone di una tabella di vettori per gli interrupt fissa.[15]

Ci sono anche quattro registri per i *segmenti* che possono essere calcolati dai registri di indice. I registri di segmento permettono alla CPU di accedere ad un megabyte di memoria in un modo particolare. Invece di fornire i byte mancanti, come nella maggior parte dei processori che supportano la segmentazione, l'8086 fa uno *shift* a sinistra di 4 bit del registro di segmento e lo somma all'indirizzo. Il risultato è che i segmenti si possono sovrapporre (parzialmente o totalmente), il che è stato considerato come un indice di cattiva progettazione da molti sviluppatori. Anche se

[13] http://it.wikipedia.org/wiki/Intel_8080
[14] http://en.wikipedia.org/wiki/Intel_8086
[15] http://it.wikipedia.org/wiki/Intel_8086

questo è un vantaggio per la programmazione in linguaggio assembly, dove il controllo sui segmenti è completo, causa invece confusione nei linguaggi che fanno molto uso dei puntatori (come ad esempio il linguaggio C). Lo schema di segmentazione dell'8086 rende difficile una rappresentazione efficiente dei puntatori ed è possibile avere due puntatori con valori diversi che puntano ad una stessa locazione di memoria. Inoltre non si può estendere facilmente per aumentare lo spazio di indirizzamento a più di un megabyte. In effetti questo è stato fatto nell'Intel 80286 cambiando radicalmente lo schema di indirizzamento.

1979

L' **8088** è un microprocessore inizialmente a 4.77MHz basato sull'8086. Ha un'architettura interna a 16 bit (come l'8086), un data bus a 8 bit (la metà dell'8086) e un address bus da 20 bit, quindi l'8088 è in grado di indirizzare 1MB di memoria. Questo processore fu utilizzato nei primi PC IBM.[16]

La funzione dell'8088 era quella di poter progettare sistemi più economici; utilizzando 8 linee in meno dell'8086 si potevano infatti costruire schede madri e chip di supporto meno complessi.

1982 – 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 tecnologia a **1.5 micron** e frequenze tra 6 e 25 Mhz.[17] Grazie alla sua capacità di gestire 16 Mbyte di memoria Ram rese possibile l'apparire delle prime arcaiche interfacce grafiche (Windows 2.0 e 2.1). Il 286, come tutti i processori Intel successivi, disponeva di una perfetta compatibilità con i programmi Dos scritti per l'8086 mentre i programmi che volevano 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. Il processore 80286 era stato progettato per applicazioni multitasking, relative alle comunicazioni (ad es. centralini telefonici PBX), processi

[16] http://en.wikipedia.org/wiki/Intel_8088
[17] http://www.aall86.altervista.org/guide/Storia_dei_processori.pdf

di controllo in tempo reale e sistemi multi-utente. Una caratteristica interessante di questo processore è che fu il primo tra quelli in architettura x86 con la possibilità di passare da modalità reale a modalità protetta, permettendo l'uso di tutta la memoria di sistema come un unico blocco e offrendo un certo grado di protezione delle zone di memoria usate dalle applicazioni. Il passaggio a tale modalità non era però reversibile: il ritorno alla modalità normale richiedeva il reset del processore. Tale limitazione fece sì che la modalità protetta non trovasse grande impiego fino all'arrivo del processore 80386, che era in grado di passare da una modalità all'altra indifferentemente.[18]

1985 – Processori di terza generazione

Il primo **80386** Dx fu realizzato nell'ottobre 1985 integrando 275.000 transistor, tecnologia a **1 micron** e frequenze di clock da 16 Mhz fino a 40 Mhz. 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 contro il massimo limite di 16 megabyte dei precedenti processori Intel a 16 Bit. 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. L'80386 fu una pietra miliare nell'evoluzione della serie di processori nata con l'Intel 8008: rispetto al suo predecessore 80286, il 80386 aveva un'architettura a 32 bit ed una unità di paginazione della memoria che rese più semplice adottare sistemi operativi dotati di gestione della memoria virtuale. L'80386 era compatibile con i vecchi processori Intel: la maggior parte delle applicazioni che giravano sui precedenti PC dotati di processori x86 (8086 e 80286) funzionavano ancora sulle macchine 80386, e perfino più velocemente. Intel produsse in seguito anche una versione a basso costo 80386Sx (Single word eXternal) con bus interno a 32 bit ed esterno a 16 bit.[19]

[18] http://it.wikipedia.org/wiki/Intel_80286
[19] http://it.wikipedia.org/wiki/Intel_80386

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, non risultava affatto più veloce del 386 a 25 Mhz. La lentezza della memoria Ram da 80 nanosecondi a cui il processore accedeva per scrivere e rileggere dati fungeva da collo di bottiglia strozzando le prestazioni. Si pensò cosi 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. 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. 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, funzionando come 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 memoria 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 si 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%.

1989 – Processori di quarta generazione

Con l'introduzione del primo Intel **80486** con clock da 16 Mhz a 50 Mhz e tecnologia da **1 a 0.6 micron** una piccola porzione di cache venne inserita all'interno dei microcircuti 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 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 si 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 nel calcolo dei numeri in virgola mobile rispetto alla accoppiata 386+387 è dovuta in parte alla cache memory 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 die 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.

Il 486 fu anche il primo processore ad adottare la tecnica del moltiplicatore interno per la frequenza di funzionamento; 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.). In questo modo si evitano i disturbi e le dispersioni di segnale sulla piastra madre dal momento che la frequenza di bus è un sottomultiplo della frequenza interna del processore.

1993 – Processori di quinta generazione

Viene realizzato il **Pentium** come successore del 486, integra fino a 3.100.000 transistor e ha frequenze operative di CPU da 60 a 300 Mhz e frequenze di bus da 50 a 60 Mhz. E’ costruito con tecnologia da 0.8 a 0.25 micron.

Le principali differenze rispetto al 486 sono[20] :

- Architettura **superscalare**: il Pentium possedeva 2 pipeline che gli permettevano di completare più di una operazione per ciclo di clock. Una pipeline, chiamata "*pipeline U*", poteva eseguire qualunque istruzione, mentre l'altra, chiamata "*pipeline V*", era in grado di eseguire solo quelle più semplici e comuni (logica cablata). L'utilizzo di più pipeline era una caratteristica delle architetture RISC; una delle tante caratteristiche che, nel tempo, sarebbero state poi implementate sulle architetture x86, dimostrando la possibilità di unire le due tecnologie e creare dei processori che possano essere definiti "ibridi".
- **Data path a 64 bi**t: questa caratteristica raddoppiava la quantità di informazioni prelevate dalla memoria in ogni operazione di fetch. È importante sottolineare però che questo aspetto non consentiva assolutamente

[20] http://it.wikipedia.org/wiki/Pentium

al Pentium di poter eseguire codice a 64 bit, dato che i suoi registri continuavano ad essere 32 bit.

- Unità **FPU** molto più veloce rispetto a quella del 486.
- Supporto per le istruzioni **MMX** (MultiMedia Extension solo per i modelli più recenti). 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. Si tratta 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 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.

1995 – Processori di sesta generazione

Il **Pentium Pro** è stato originariamente sviluppato e prodotto come sostituto del Pentium. Realizzato con tecnologia BiCMOS da da **0.5 a 0.35** micron conteneva fino a **5.5 milioni** di transistor con CPU clock da 150 a 200 Mhz e FSB clock da 60 a 66 mhz.[21]

Le principali caratteristiche del Pentium Pro sono :

• **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

[21] http://en.wikipedia.org/wiki/Pentium_Pro

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. In questo modo la CPU poteva accedere contemporaneamente ai dati in cache e ai dati in ram riducendo enormemente i tradizionali rallentamenti dovuti agli accessi contemporanei e ciò era evidente soprattutto nelle operazioni di I/O. In un ambiente multiprocessore la cache integrata del Pentium Pro faceva la differenza rispetto ai processori che condividevano una unica cache comune. Questo rendeva il Pentium Pro costosissimo da produrre ma tale caratteristica, abbandonata con il Pentium II sarà reintrodotta in seguito nel Celeron e da li in tutti i processori successivi.

• **Superpipeline**: è 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. Pipeline profonde possono sì garantire alte frequenze operative, ma in caso di errori nella previsione per l'elaborazione anticipata del codice, è necessario svuotare completamente la catena di montaggio per riempirla con i dati corretti. Ciò significa perdere preziosi cicli di elaborazione, che diminuiscono le prestazioni generali del sistema. La pipeline quindi consente l'elaborazione di nuovi dati senza necessità di attendere che i dati precedentemente inviati termino la loro elaborazione. Con questo metodo i processori centrali, o quelli della scheda grafica, possono ricevere dati mentre ne elaborano altri, migliorando l'efficienza del sistema.

• **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 implementava al suo interno tre 486 operanti in parallelo.

• **Esecuzione fuori ordine (Out of order execution)**: Nel Pentium, 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). 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.

L'architettura P6 del Pentium Pro rimarrà valida anche per i successivi processori quali il **Pentium II** (**1997**, 400 Mhz , 7.5 milioni di transistor a 0.35 micron) e il **Pentium III** (**1999**, 400 Mhz a 1.4 Ghz, 0.25 a 0.13 micron) e per il **Pentium M** che verrà progettato come una versione profondamente modificata del Pentium III (per poter essere utilizzato sui pc portatili). È ottimizzato per un basso consumo, una caratteristica fondamentale per estendere la durata della batteria dei computer portatili.

2000 – Processori di settima generazione (Netburst architecture)

L'architettura NetBurst è stata sviluppata quando la strada maestra per aumentare le prestazioni sembrava l'innalzamento della frequenza operativa. Si trattava infatti di un'architettura nata per spingere il processore fino a frequenze di 10 GHz e, a questo scopo, dotata di pipeline molto lunghe che, nella sua ultima realizzazione, sono arrivate fino a 31 stadi. Pipeline così lunghe subiscono penalizzazioni elevatissime in caso di salti non predetti correttamente o in caso di istruzioni che devono stallare per la mancanza di qualche risorsa. Per ridurre al minimo il problema NetBurst implementava praticamente tutte le tecniche disponibili per ridurre le condizioni di stallo delle pipeline e integrava più pipeline per sfruttare il parallelismo del codice.[22]

[22] http://it.wikipedia.org/wiki/NetBurst

Il **Pentium 4** è il capostipite di questa architettura con una frequenza di clock da 1.3 a 3.8 Ghz, tecnologia da 0.18 a 0.065 micron, FSB da 400 a 1600 Mhz.

Il nuovo processore non migliorava il design P6 né nel calcolo intero, né in virgola mobile, generalmente considerati i fattori chiave nelle prestazioni di un processore. Furono sacrificate le prestazioni nel singolo ciclo di clock per guadagnare su due fronti: nella massima frequenza raggiungibile, e nelle prestazioni sfruttando le nuove librerie SSE2 che andavano ad aggiungersi alle precedenti SSE (**S**treaming **S**IMD **E**xtensions) ed MMX.

Il Pentium 4 "svolge meno lavoro" in ogni ciclo di clock rispetto ad altre CPU (come ad esempio i vecchi Pentium III), ma l'obiettivo iniziale di sacrificare le prestazioni sul singolo ciclo di clock era bilanciato dalla possibilità di aumentare molto velocemente la frequenza di funzionamento, caratteristica che portava comunque a ottime prestazioni. L'aumento della velocità di clock trovò il suo limite quando gli ingegneri dovettero affrontare problemi insolubili di eccessiva produzione di calore. Quest'ultimo elemento è il risultato della presenza di correnti di dispersione (leakage) all'interno del chip legate all'inefficienza dei circuiti e degli stessi transistor costruiti con il silicio. Parte dell'energia fornita, infatti, viene dispersa sotto forma di calore, calore che con la densità di impacchettamento raggiunte dai chip attuali ne comprometterebbe il funzionamento portando il chip ad una temperatura maggiore della soglia dei 90° tollerata. Il massimo valore di clock raggiunto fu quello di 3.8 Ghz nel 2004 ma con un TDP (thermal design power) di 115 W per chip che portò all'abbandono di tale microarchitettura in favore della successiva microarchitettura **Core**.

Caratteristiche tecniche dell'architettura Netburst

Hyper Pipelined Technology

Questo è il nome che Intel ha scelto per la pipeline a venti stadi prevista dalla prima generazione dell'architettura NetBurst. Questo è un aumento significativo, paragonato ai soli 10 stadi della pipeline del Pentium III. Una pipeline così lunga ha comunque degli svantaggi, in particolare un ridotto IPC (Istruzioni Per Ciclo),

mitigato però dalla possibilità di aumentare la velocità di clock. Un altro svantaggio è dato dal gran numero di stadi che devono essere ripercorsi nel caso in cui l'algoritmo di branch prediction faccia un errore (cache miss). Per limitare i danni (miss penalty) dovuti a tali inevitabili problemi, Intel ha introdotto le tecnologie "Rapid Execution Engine" e "Execution Trace Cache" e ha raffinato l'algoritmo di branching, migliorando notevolmente la hit rate.

Rapid Execution Engine

Intel ha aggiunto due unità per le operazioni con gli interi nella ALU rispetto all'architettura P6. Le aggiunte sono un addizionatore per interi e una unità di calcolo per gli indirizzi. Ma la novità più importante introdotta da questa tecnologia è la velocità di clock della ALU, che opera al doppio della velocità di clock del core. Questo vuol dire che in una CPU a 3 GHz la ALU opera a 6 GHz. Queste migliorie combattono il calo di IPC e migliorano di molto le prestazioni della CPU nei calcoli sugli interi. Lo svantaggio è che alcune istruzioni sono più lente, come ad esempio lo *shift*, dovuto alla mancanza di un barrel shifter (circuito digitale che può shiftare un *data word* di un numero predefinito di bit in un solo ciclo di clock) , che era incorporato in ogni CPU dal 80386.

Execution Trace Cache

All'interno della cache L2 della CPU Intel ha incorporato una **Execution Trace Cache**. Questa cache memorizza le micro operazioni dopo lo stadio di decode, cosicché quando deve passare a una nuova operazione, invece di dover eseguire di nuovo il fetching e il decoding dell'istruzione, la CPU può accedere direttamente alle micro-operazioni dalla trace cache, risparmiando una notevole quantità di tempo. Inoltre le micro-operazioni sono mantenute nella cache secondo l'ordine di esecuzione predetto algoritmicamente, il che significa che quando la CPU recupera le istruzioni dalla cache, esse sono già presenti nell'ordine corretto.

Capitolo 3. Tecnologie di produzione

§ 1 – Evoluzione della tecnologia del processo di fabbricazione

L'evoluzione dei microprocessori non dipende esclusivamente dalla progettazione di nuove architetture sempre più efficienti, ma anche dall'evoluzione dei processi produttivi. Negli anni le dimensioni dei transistor che compongono una CPU sono drasticamente diminuite e questo ha consentito di aumentarne notevolmente il numero all'interno di un unico processore e al contempo di poter impiegare clock via via crescenti. Per comprendere l'entità di questo fenomeno, è sufficiente vedere l'evoluzione del processo produttivo dal primo processore Pentium (una delle pietre miliari nella storia del CPU) presentato nel 1993 e costruito a 800 nm (nanometri); integrava 3,1 milioni di transistor e funzionava a 60 MHz. Nel 2007 viene presentato il Core 2 Extreme Yorkfield costruito a 45 nm, integrante 820 milioni di transistor e con una frequenza di 3 GHz.[23]

Un processo produttivo più avanzato permette di diminuire i consumi e in genere anche di poter aumentare il clock di funzionamento; a questo si aggiunge il fatto che avere processori "più piccoli" significa poterne produrre di più sullo stesso wafer di silicio, con ovvie ripercussioni in termini di costi finali dei prodotti e soprattutto di margini di guadagno da parte del produttore. È da sottolineare però, come lo sviluppo di un nuovo processo produttivo sia una difficoltà di tipo tecnico di competenza quasi esclusiva ingegneria dei materiali e dell'ingegneria meccanica, più che per quella elettronica. Dato che lo scopo dei produttori di microprocessori è quello di aumentare continuamente le cosiddette "*Prestazioni per Watt*", ovvero migliorare sempre di più l'efficienza di un microprocessore per sfruttarne al massimo il proprio potenziale, diventa necessario sviluppare parallelamente ad un nuovo processo produttivo anche le architetture delle CPU.

[23] http://it.wikipedia.org/wiki/Intel_%28approccio_ciclo_evolutivo_cpu%29

La filosofia adottata da Intel per procedere nello sviluppo della tecnologia è quella del modello **tick-tock**. [24]

Con la fase **tick** si intende una diminuzione delle dimensioni del precedente processo tecnologico di fabbricazione, aumentando la densità dei transistor sul die, applicato sulla esistente microarchitettura.

Una volta che la nuova tecnologia è funzionante e a regime si passa alla fase **tock** che consiste nell'applicare il nuovo processo di fabbricazione ad una nuova microarchitettura.

Dall'inizio del 2006, la strategia di Intel è quella di introdurre una nuova architettura ogni 2 anni, e precisamente ogni anno pari (fase tock) e introdurre un nuovo processo di fabbricazione negli anni dispari (fase tick).

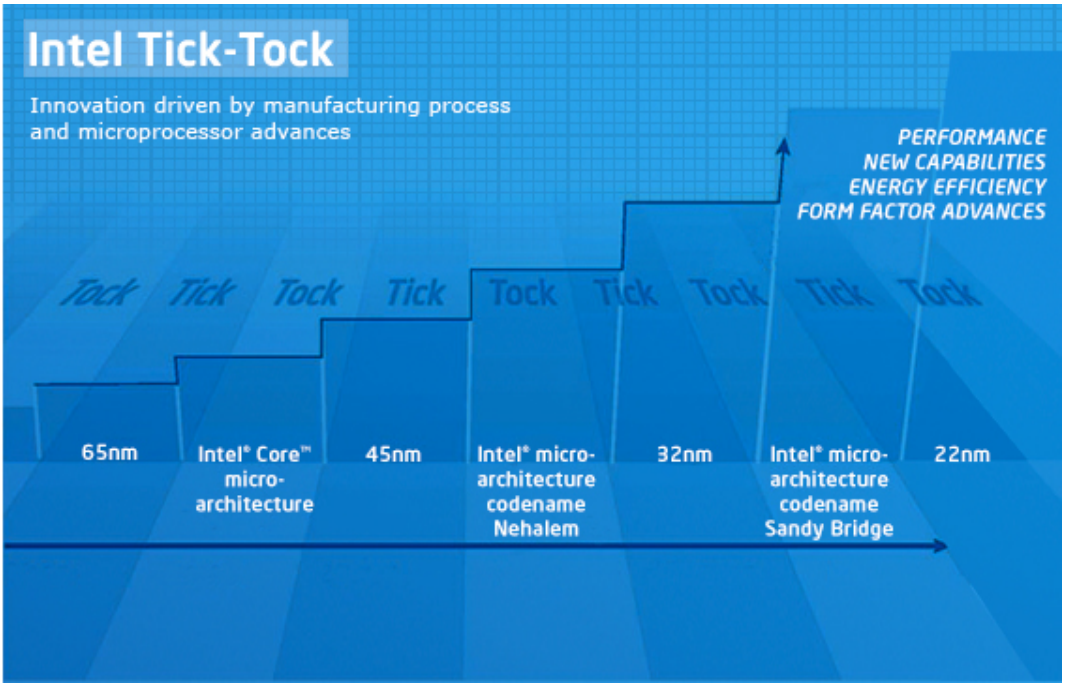

Figura 4 Modello tick-tock per l'evoluzione dei microprocessori Intel [25]

[24] http://en.wikipedia.org/wiki/Intel_Tick-Tock#cite_note-6
[25] http://www.intel.com/technology/tick-tock/

Architectural change		Codename	Fabrication process	Release date	Processors			
					Enthusiast	Desktop	Mobile	Marketing names
Tick	Die shrink	Presler, Cedar Mill, Yonah	65 nm	January 5, 2006	Presler	Cedar Mill	Yonah	▪ Core ▪ Pentium 4 ▪ Pentium D ▪ Pentium M ▪ Pentium Dual-Core ▪ Celeron
Tock	New microarchitecture	Core		July 27, 2006	Kentsfield	Conroe	Merom	▪ Core 2 ▪ Pentium Dual-Core ▪ Pentium ▪ Celeron Dual-Core ▪ Celeron ▪ Celeron M
Tick	Die shrink	Penryn	45 nm	November 11, 2007	Yorkfield	Wolfdale	Penryn	
Tock	New microarchitecture	Nehalem		November 17, 2008	Bloomfield	Lynnfield	Clarksfield	▪ Core i3 ▪ Core i5 ▪ Core i7 ▪ Pentium ▪ Celeron
Tick	Die shrink	Westmere	32 nm	January 4, 2010	Gulftown	Clarkdale	Arrandale	
Tock	New microarchitecture	Sandy Bridge		January 9, 2011	Sandy Bridge-EX	Sandy Bridge-DT	Sandy Bridge-NB	
Tick	Die shrink	Ivy Bridge	22 nm	2012				
Tock	New microarchitecture	Haswell		2013				
Tick	Die shrink	Broadwell	14 nm	2014				
Tock	New microarchitecture	Skylake		2015				
Tick	Die shrink	Skymont	10 nm	2016				
Tock	New microarchitecture			2017				

Figura 5 Mappa temporale del modello tick-tock per i microprocessori Intel

§ 1.1 – Tecnologia 65nm

Secondo la Legge di Moore, il numero di transistor contenuti un chip raddoppia quasi ogni due anni, con un conseguente aumento delle caratteristiche e delle prestazioni e una riduzione del costo a transistor.[26] La riduzione delle dimensioni dei transistor comporta lo sviluppo di problemi di consumo di energia e dissipazione del calore.

La tecnologia "**silicon strained**" di Intel, implementata per la prima volta nel processo di fabbricazione a 90 nm, è stata ulteriormente migliorata in quello a 65 nm.

Questa tecnologia prevede la modifica della struttura reticolare del silicio del canale, stirandolo o comprimendolo per ottenere una maggiore mobilità di elettroni e lacune e di conseguenza una maggiore velocità di commutazione dei transistor Nmos nel caso di stiramento e dei transistor Pmos nel caso di compressione.[27]

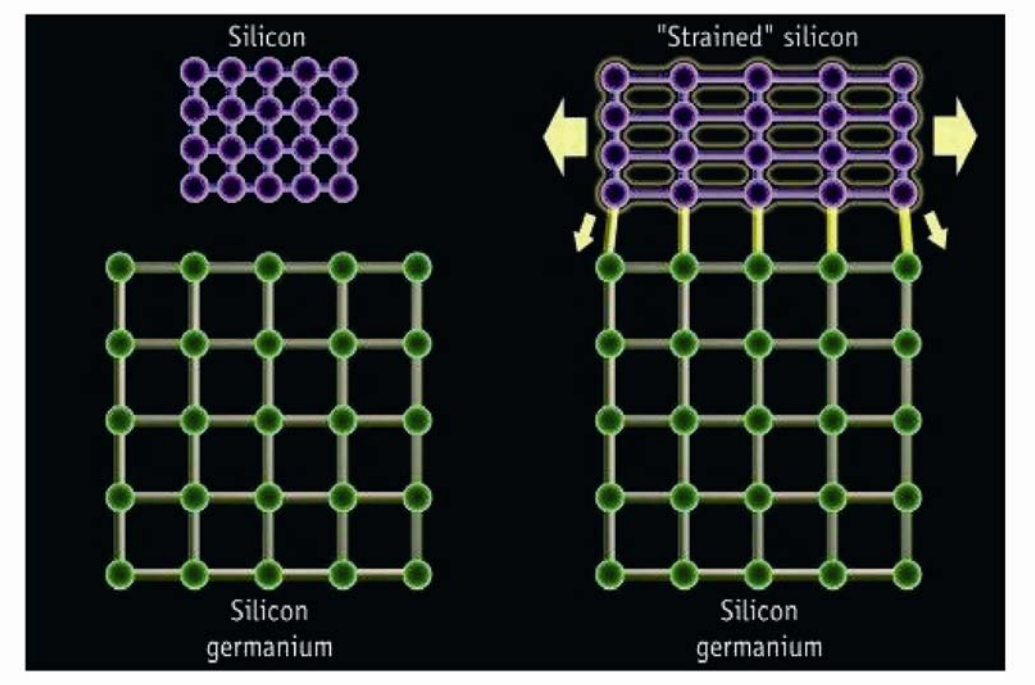

Figura 6 Tecnologia Silicon Strained

[26] http://www.intel.com/cd/corporate/techtrends/emea/ita/209823.htm
[27] http://www.intel.com/pressroom/kits/45nm/IntelHigh-K_metal_gate_glossary_FINAL.pdf

La seconda generazione di silicio strained aumenta le prestazioni dei transistor del 10-15% senza incrementare la dispersione di corrente, ma riducendola di quattro volte a prestazioni costanti rispetto ai transistor a 90 nm. Di conseguenza, i transistor realizzati con il processo a 65 nm di Intel prevedono prestazioni più elevate senza aumenti significativi della corrente di dispersione (una dispersione di corrente maggiore comporta una generazione di calore maggiore).

§ 1.2 – Tecnologia 45nm

L'introduzione della tecnologia **Hi-k metal gate** (HK+MG) nel processo a 45 nm permette di ottenere il raddoppio nella densità dei transistor, una velocità di commutazione superiore del 20%, una riduzione della corrente di dispersione source-drain di 5 volte e una riduzione della corrente di dispersione attraverso il dielettrico del gate di 10 volte.[28]

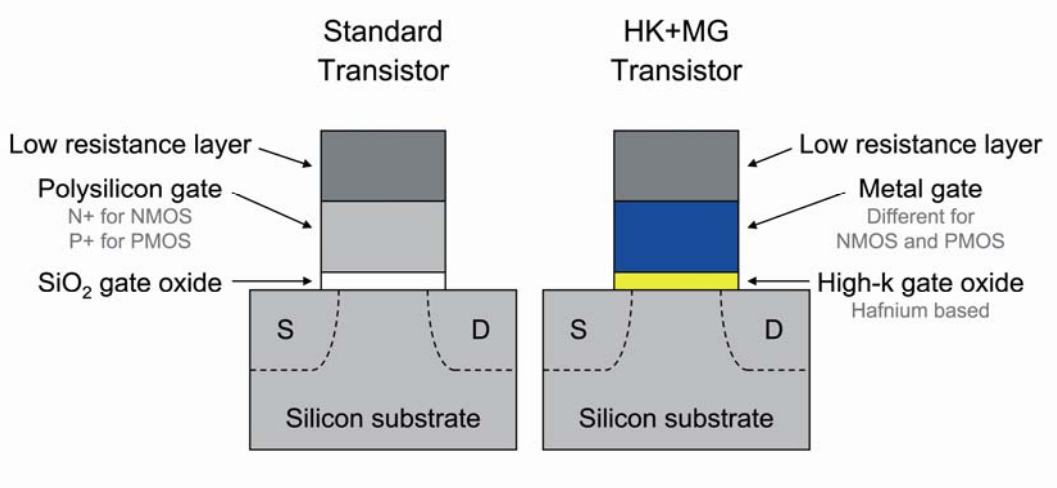

Figura 7 Confronto tra transistor Standard e High-k Metal gate

Il principio di funzionamento dei transistor di questo tipo prevede che nello stato "on" la corrente che fluisce tra il source e il drain sia il più alta possibile mentre nello stato di "off" sia la più bassa possibile.

La diminuzione dello spessore del dielettrico del gate è necessaria per aumentare l'effetto di campo e serve ad aumentare le correnti di "on" e diminuire quelle di "off".

[28] download.intel.com/pressroom/kits/45nm/Press45nm107_FINAL.pdf

Il problema nasce quando nel processo di miniaturizzazione si diminuisce troppo lo spessore del dielettrico perché cominciano a presentarsi correnti di dispersione che fluiscono attraverso il dielettrico peggiorando la situazione dal momento che al di sotto del gate di silicio policristallino si forma una zona priva di cariche (zona di deplezione) che contribuisce ad aumentare lo spessore del dielettrico avendo come effetto la diminuzione delle correnti di "on" e l'aumento di quelle di "off".

L'utilizzo di un gate basato su di un metallo (differente per transistor N o Pmos) permette di ottenere un significativo aumento dell'effetto di campo a parità di tensione utilizzata, necessario ad attivare il transistor.

La seconda grande innovazione consiste nell'eliminare anche dalla parte isolante del gate il silicio, sostituendolo con un altro materiale dielettrico. La lettera K solitamente indica in ingegneria la capacità di un materiale di trattenere la carica elettrica al proprio interno, come farebbe una spugna con dell'acqua, e viene chiamata capacitanza. Il materiale utilizzato è basato sull'elemento chimico Afnio, ed è caratterizzato da ottime doti isolanti e da una elevatissima costante K, da cui prende il nome la tecnologia High-K. La buona capacità isolante riduce i problemi dovuti alle correnti parassite, e grazie all'ottima capacità di trattenere la carica elettrica permette un notevole aumento dell'effetto di campo a parità di tensione utilizzata.

In pratica la tecnologia HK+MG aumenta la velocità di commutazione del 20% e riduce di un fattore da 5 a 10 le correnti di dispersione.

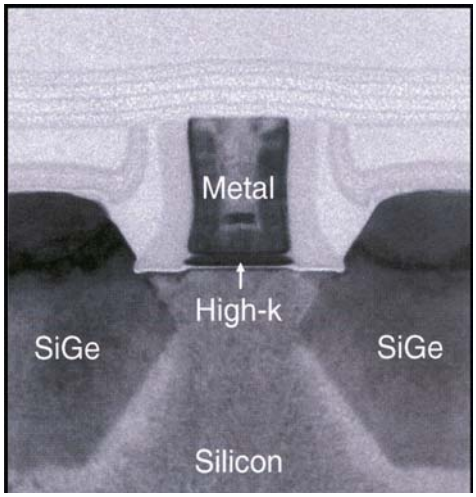

Figura 8 Transistor High-k + Metal gate (HK+MG)

§ 1.3 – Tecnologia 32nm

Il fondamento del processo di costruzione a 32 nm è la seconda generazione del transistor in tecnologia HK+MG che ha visto la sua nascita con la tecnologia a 45 nm.[29] I miglioramenti rispetto al transistor della prima generazione sono molteplici.
Lo spessore del dielettrico passa da 1.0 nm da 0.9 nm mentre la lunghezza del gate è stata ridotta a 30nm.
Il diminuito spessore del dielettrico e la ridotta lunghezza del gate contribuiscono ad un aumento della performance superiore al 22%.
La corrente di dispersione viene ridotta di un fattore 5 per i transistor Nmos e di un fattore 10 per quelli Pmos.
Il gate pitch (la distanza minima tra due gate di transistor contigui), diminuendo di un fattore 0.7x ogni 2 anni, raggiunge i 112.5 nm.

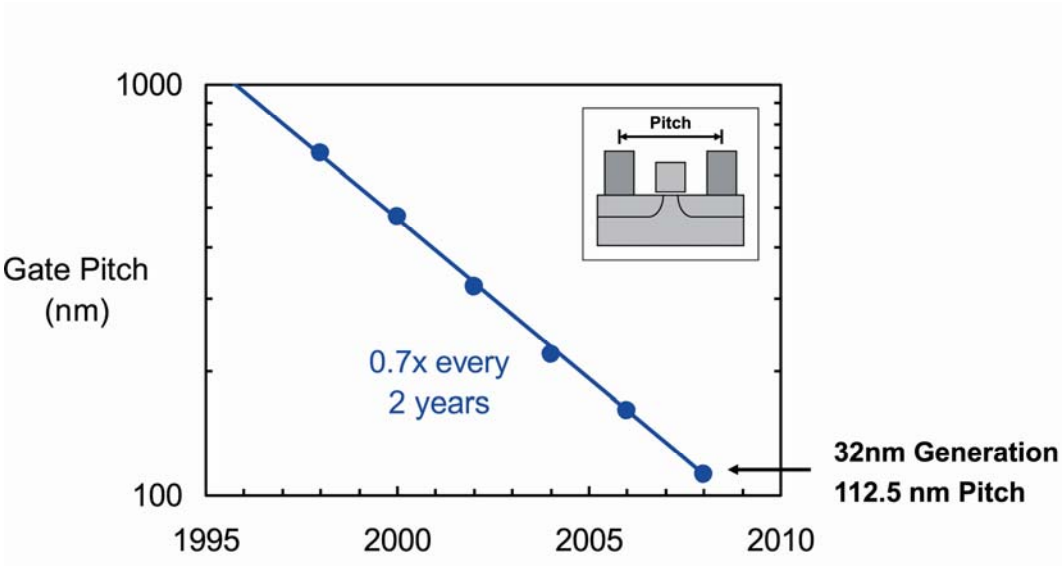

Figura 9 Grafico della diminuzione di dimensioni del gate pitch.

[29] White paper Introduction to Intel's 32 nm process technology

§ 1.4 – Tecnologia 22nm

L'introduzione del processo a 22 nm segna un momento storico nell'evoluzione dei transistor alla pari con quello dello stop alla corsa all'innalzamento di frequenza dei microprocessori da parte dei produttori di cpu[30]. Oggi per i transistor vale lo stesso discorso: è finita l'era della miniaturizzazione a oltranza e inizia quella geometria di costruzione, un nuovo stadio evolutivo necessario per rispettare la legge di Moore che non si esclude debba essere riformulata nei prossimi anni.

A partire dal processo produttivo a 45 nm Intel ha cominciato a modificare gli elementi stessi di costruzione del transistor: dai soli elementi semiconduttori drogati (fondamentalmente silicio) si passò ad un modello con uno strato altamente isolante a base di afnio e a un gate metallico in grado di condurre meglio la corrente elettrica. Nonostante questa nuova tecnologia la sua applicazione ai 22 nm si scontra con il fatto che il gate poggia su un canale largo solo 22 nm e la superficie di contatto è estremamente ridotta , rendendo molto difficile la creazione di un canale adatto allo spostamento delle cariche elettriche e al funzionamento pratico di un transistor.

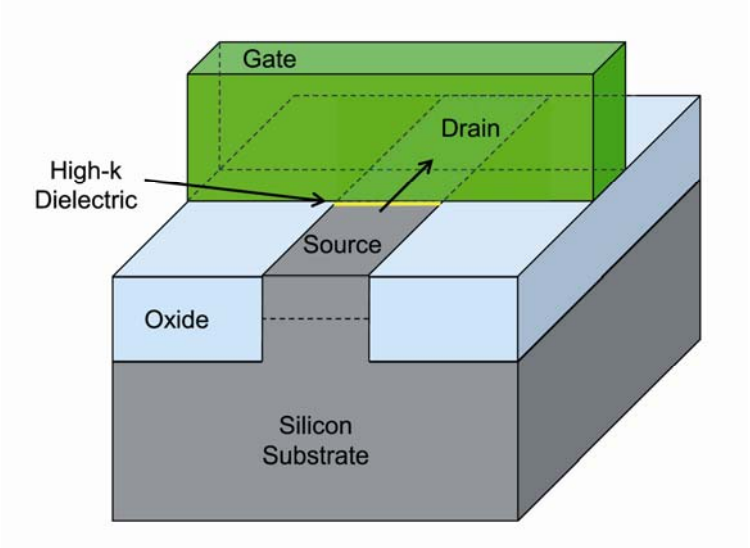

Figura 10 Transistor 2D [31]

[30] Pc Professionale N. 244 (Luglio 2011) p. 92 - 99
[31] M.Bohr,M.Kaizad "Intel's Revolutionary 22 nm Transistor Technology"

Intel ha superato questo problema elevando nella pratica il canale del gate tra source e drain all'interno del gate stesso, rendendolo di conseguenza circondato da tre lati invece di uno solo.

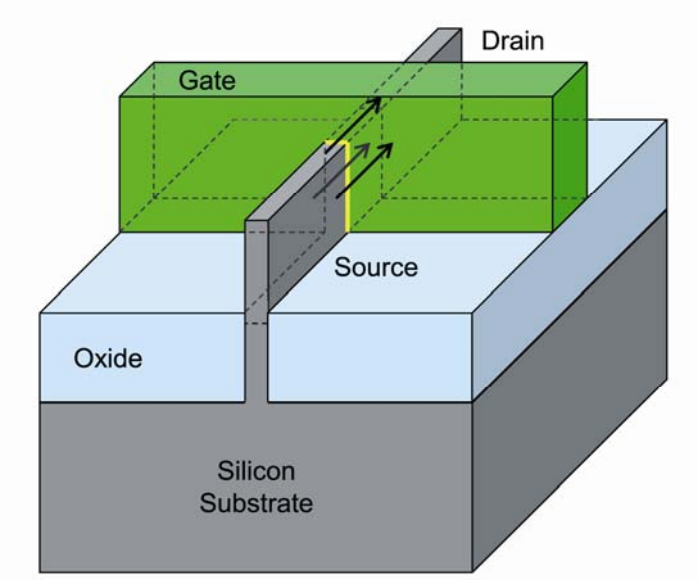

Figura 11 Transistor tri-gate 3D

La superficie esposta, potendo elevare anche di parecchi nm questo canale (detto anche aletta o **fin**), risulta nel complesso estremamente superiore rispetto alla sola superficie orizzontale esposta in precedenza. Dal momento che lo scopo ultimo del transistor è ottenere il passaggio di corrente tra source e drain quando sul gate c'è tensione e impedirlo quando quest'ultima è assente, ci si rende conto come una superficie esposta maggiore permette di ottenere questo risultato utilizzando tensioni nel gate inferiori (dal momento che il campo elettromagnetico viene creato sia da destra che da sinistra che dall'alto) e distanze minori tra source e drain a tutto vantaggio della dimensione planare del transistor che ora si sviluppa anche in altezza.

Con questa tecnica Intel ha risolto uno dei problemi fondamentali dei transistor moderni: **la regione di svuotamento** (depletion region) al di sotto del canale del gate (inversion layer).

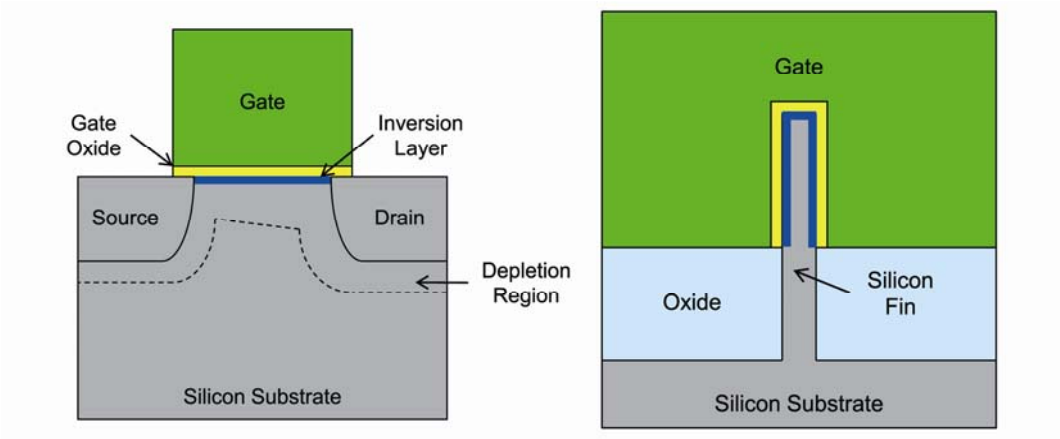

Figura 12 Il substrato di silicio esercita un'influenza sullo strato di inversione che è lo strato in cui fluisce la corrente source-drain. Dal momento che le caratteristiche di turn-off sono degradate dalla regione di svuotamento, quanto più questa è sottile rispetto allo strato di inversione, tanto meno si presentano le correnti parassite che degradano lo stato di turn-off.

Nella pratica della realizzazione a 22 nm Intel utilizza un numero elevato di alette parallele massimizzando gli effetti del progetto 3D tri-gate distribuendo tensioni e correnti tra i vari elementi.

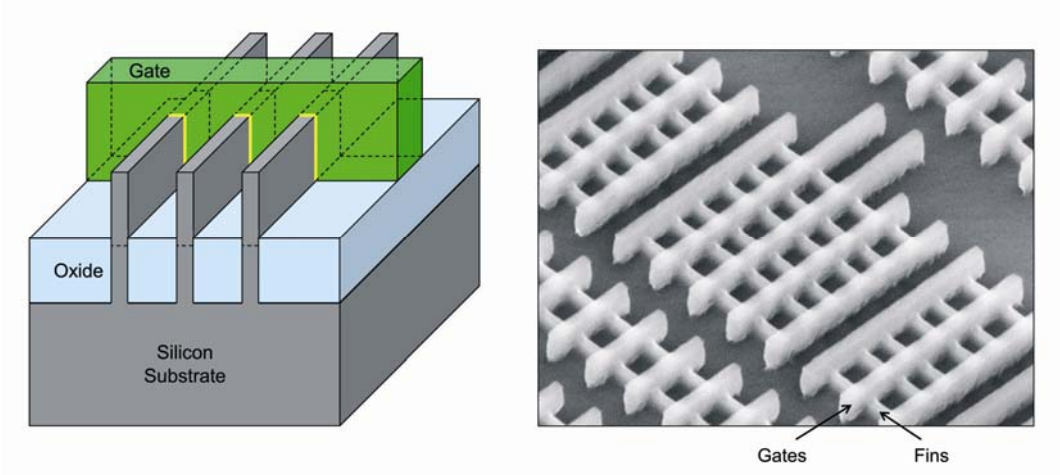

Figura 13 Transistor tri-gate multiplo

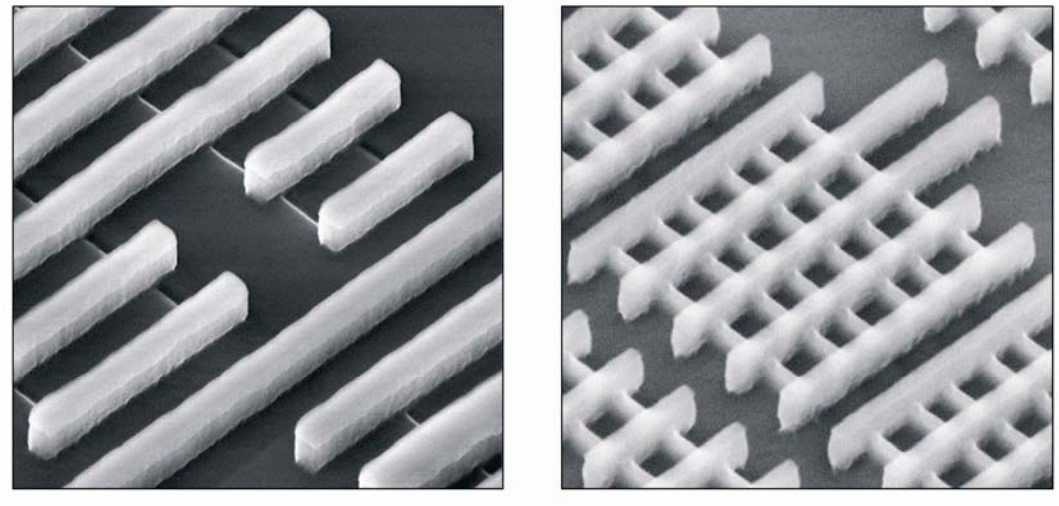

Figura 14 Transistors planari a 32 nm Transistors tri-gate multipli a 22 nm

I vantaggi operativi del transistor tri-gate sono molteplici.

Con questa architettura costruttiva è possibile ridurre la corrente parassita tra source e drain in stato “off” di un fattore 10 oppure, mantenendo lo stesso livello di correnti parassite dei transistor planari, operare con una tensione inferiore di circa 0,15 V.

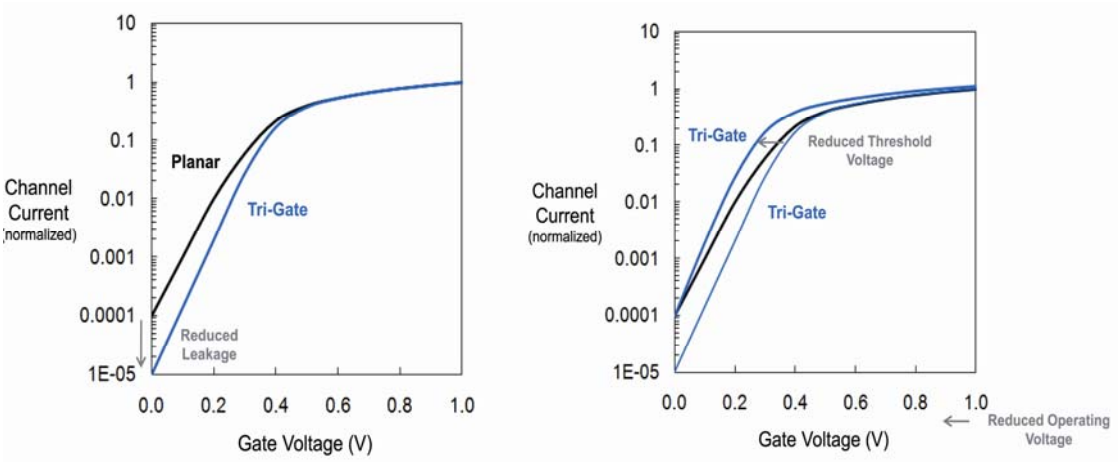

Figura 15 Riduzione delle correnti di dispersione o possibile riduzione di tensione di gate

Per la velocità di commutazione abbiamo un miglioramento tra il 18% e il 37% oppure, mantenendo inalterate le prestazioni, è possibile operare con 0,2 V in meno con enormi risparmi nei consumi e nel calore prodotto.

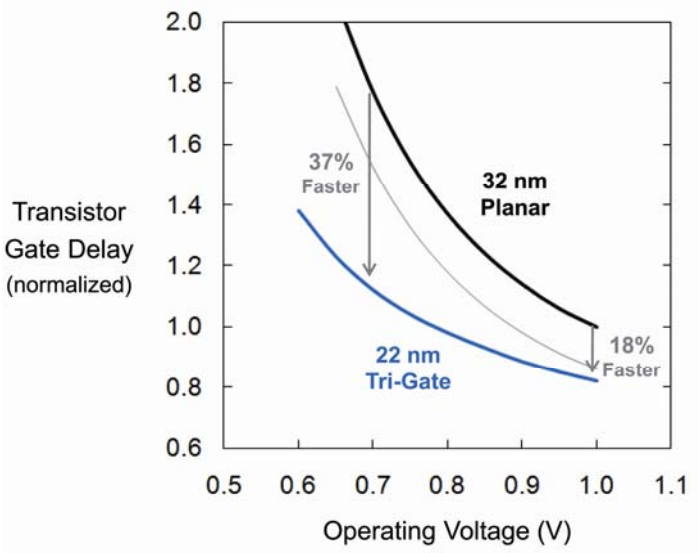

Figura 16 Velocità di commutazione dei nuovi transistor 3D tri-gate

Intel riporta che un transistor 3D a 22 nm può arrivare a raggiungere frequenze di commutazione superiori a 100 Ghz ![32]

[32] download.intel.com/newsroom/kits/22nm/pdfs/22nm_Fun_Facts.pdf

Capitolo 4. Microarchitetture attuali

§ 1 – Microarchitettura recenti

§ 1.1 Processori di ottava generazione (2006 – Core 2 65nm)

E' una architettura multi processore basata su una versione ottimizzata del Pentium M e quindi su un processore della sesta generazione iniziata dal Pentium Pro. Intel abbandona quindi l'architettura Netburst della precedente generazione di processori a causa dell'alto consumo, del calore sviluppato all'aumentare della frequenza di clock e di altre problematiche come l'inefficienza della pipeline (dovuta essenzialmente alla sua lunghezza).[33] Abbiamo modelli single e dual core su singolo die e modelli quad core su due die separati nello stesso package. Fabbricato con tecnologia iniziale a 65nm poteva contenere fino a 500 milioni di transistor a seconda dei core presenti.

La nuova architettura "Core" vede l'introduzione di molte innovazioni rispetto al passato.[34]

1) Wide Dynamic Execution

Attraverso questa tecnologia è possibile eseguire più istruzioni per ciclo di clock rispetto a quanto era possibile nei processori basati sulle architetture precedenti. Ogni core può ora completare 4 istruzioni contemporaneamente, contro le 3 consentite da NetBurst. La lunghezza della pipeline si è accorciata molto rispetto a quella impiegata precedentemente, infatti il primo processore Netburst, ovvero il Pentium 4 Willamette, aveva una pipeline a 20 stadi che erano poi saliti a ben 31, nell'ultima evoluzione del Pentium 4, il core Prescott. L'architettura "Core" invece riprende la pipeline a 14 stadi. È necessario sottolineare come una pipeline più corta sia meno vulnerabile ai salti nella successione di istruzioni e nella lettura di dati dalla memoria RAM, anche se rende più difficile raggiungere frequenze di clock elevate: si tratta "solo" di trovare il giusto bilanciamento.

[33] http://en.wikipedia.org/wiki/Intel_Core_%28microarchitecture%29
[34] http://download.intel.com/technology/architecture/new_architecture_06.pdf

All'interno di questa tecnologia ne trova posto anche un'altra chiamata "***MacroFusion***" che consente di unire tra loro alcune istruzioni per ottenere un'elaborazione più veloce e diminuire il consumo dal momento che la cpu esegue una sola micro-operazione invece di due. Alcune coppie di istruzioni tipiche come ad esempio CMP (compare) e il successivo salto condizionato (JNE) vengono combinate in una singola istruzione interna (micro-op) durante la decodifica. Quindi due istruzioni di programma possono essere eseguite come un'unica micro-op riducendo in questo modo il lavoro del processore e aumentando di conseguenza il numero di istruzioni che possono essere eseguite in ogni ciclo di clock.

2) Advanced Smart Cache

La cache L2 di un processore dual core viene condivisa da ciascun core. I vantaggi di tale tecnologia sono molteplici, infatti se da una parte viene minimizzato il traffico di dati sul bus rispetto ad una soluzione dual core a 2 cache separate, dall'altra consente ad un core di utilizzare l'intera cache nel caso in cui l'altro core fosse al momento inattivo, cosa che può facilmente accadere con tutte quelle applicazioni che non sono in grado di sfruttare la presenza di più di un core in un sistema. Un altro vantaggio che deriva da questo tipo di implementazione è l'impossibilità che uno stesso dato possa essere duplicato nella cache L2, cosa che poteva accadere con i Pentium D dove le cache, essendo separate per ciascun core, potevano contenere dati replicati. Quindi ogni core può dinamicamente utilizzare fino al 100% della cache complessiva a disposizione dei vari core presenti sul die riducendo le cache misses e aumentando le prestazioni.

3) Smart Memory Access

Attraverso altri miglioramenti vari, si è potuto ottenere un generale abbassamento delle latenze di accesso alla memoria RAM. E' importante notare che il codice del software per l'architettura x86 contiene circa il 38% di istruzioni di Load e Store[35]. Generalmente le istruzioni di Load sono il doppio di quelle di Store.

[35] White Paper Inside Intel®Core™ Microarchitecture and Smart Memory Access

Per prevenire l'inconsistenza dei dati le istruzioni sono normalmente eseguite come appaiono nel programma nel senso che se una istruzione specifica uno "Store" ad un indirizzo particolare e quindi un "Load" dallo stesso indirizzo queste due istruzioni devono essere eseguite nell'ordine stabilito. Ma cosa accade se gran parte delle istruzioni non hanno questo tipo di dipendenza ? Come può questa non-dipendenza essere utilizzata per migliorare l'efficienza del processore e la velocità di esecuzione delle istruzioni ?

Nell'implementazione dell'architettura Core Intel ha trovato il metodo per identificare le false dipendenze tra le istruzioni di Store e Load tramite la tecnica di Memory Disambiguation (intesa come rimozione dell'ambiguità).Tramite questa tecnica la microarchitettura Core è in grado di risolvere molti casi di non-dipendenza tra Store e Load in modo da poter procedere con l'esecuzione out-of-order delle istruzioni di Load. In questo modo limitando al massimo la possibilità di ambiguità della memoria si sfrutta al meglio la pipeline ed si evitano svuotamenti della stessa a causa di dati non ancora disponibili. Si tratta di una innovazione che va a risolvere un vero tallone d'Achille della precedente architettura NetBurst.

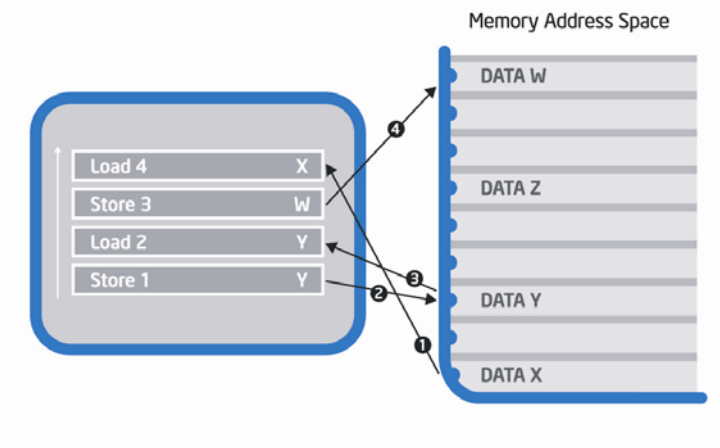

Figura 17 In questo esempio di memory disambiguation, i numeri cerchiati indicano l'ordine cronologico di esecuzione e la freccia all'estrema sinistra indica la sequenza di programma. Come si può vedere, Load 2 non può essere anticipato dal momento che deve aspettare l'esecuzione di Store1 affinchè la variabile Y abbia il valore corretto. Tuttavia, il memory disambiguation predictor può riconoscere che Load 4 non dipende dall'ordine delle altre istruzioni che appaiono e può essere eseguito prima senza aspettare l'esecuzione né di Store 3 né di Store1. Nell'eseguire Load 4 alcuni cicli in anticipo, la CPU ha subito i dati richiesti da qualsiasi istruzione seguente che abbia bisogno del valore di X, riducendo i tempi di latenza della memoria e introducendo quindi un più alto grado di parallelismo a livello di esecuzione delle istruzioni.

4) Advanced Digital Media Boost

Questa caratteristica aumenta le prestazioni quando il processore esegue istruzioni di tipo SSE (Streaming SIMD Extension). Sono queste istruzioni a 128 bit intere e floating point a doppia precisione tipiche dell'utilizzo di software che trattano video, immagini, elaborazione di foto, criptaggio e applicazioni scientifiche in genere. Nelle precedenti generazioni di processori le istruzioni a 128 bit di tipo SSE, SSE2, SSE3 venivano eseguite in due cicli di clock, prima i 64 bit inferiori e quindi i rimanenti 64 bit. Nella generazione Core, grazie all'utilizzo di un datapath interno a 128 bit l'istruzione è eseguita in un solo ciclo di clock, 128 bit alla volta, raddoppiando in pratica la sua velocità di esecuzione.

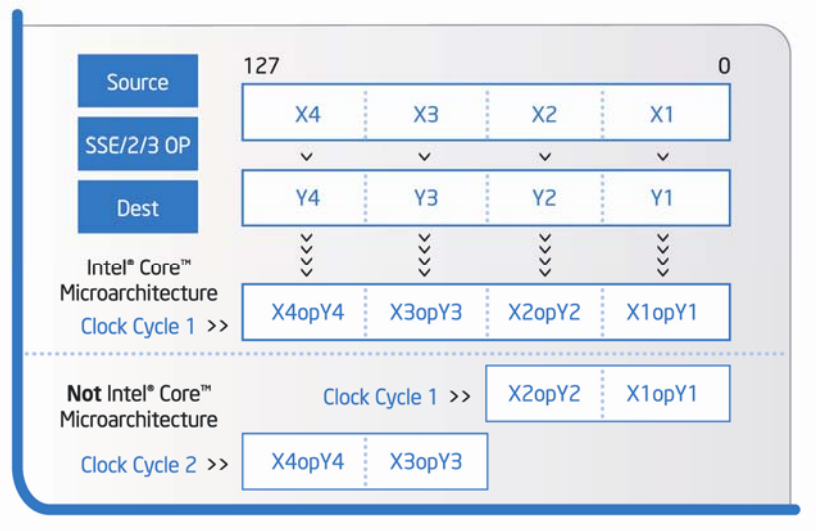

Figura 18 Esempio di esecuzione di una istruzione SSE a 128 bit. Raddoppio della velocità di esecuzione e diminuzione di consumo energetico rispetto ad una microarchitettura precedente.

5) Intelligent Power Capability

Questa caratteristica abilita la CPU a porre in stato di minima energia i core che non sono usati ma non solo, può arrivare a "spegnere" specifiche unità logiche all'interno di ogni core fino a gestire le singole linee dei bus interni abilitando solo le linee di bus necessarie all'istruzione in esame.[36]

[36] http://www.hardwaresecrets.com/article/Inside-Intel-Core-Microarchitecture/313

Molti dei bus interni alla CPU sono dimensionati per il "caso peggiore" ossia l'istruzione più grande che nell'architettura x86 è lunga 15 Byte (480 bit) e risulta quindi conveniente da un punto di vista energetico abilitare solo le linee di bus che servono effettivamente in base all'istruzione in esecuzione.
Ad esempio, l'istruzione mov eax (dato a 32 bit) che memorizza il dato a 32 bit all'interno del registro EAX della CPU internamente è considerata essere una istruzione a 40 bit (8 bit di opcode più i 32 bit del dato) e quindi abbiamo 440 linee di bus che possono essere temporaneamente disabilitate per non consumare energia.
Questa tecnologia, mutuata dal mondo mobile, è stata ulteriormente affinata e resa più granulare: lo spegnimento e il ripristino avviene in tempi molto più rapidi e con minori dispendi energetici. Per esempio, nell'architettura Core sono supportati 5 C-state (C0-C4), stati di idle progressivi in cui il processore passa a livelli sempre più a basso consumo: si passa dai 35 W dello stato C0 ai 1,2 W di quello C4.[37]

§ 1.1.1 - Evoluzione di Core (2008 - Penryn 45nm)

Nell'evoluzione a 45 nm della microarchitettura Core (architettura **Penryn**) viene introdotto un nuovo stato di idle, che prende il nome di **Deep Power Down Technology**. Rispetto allo stato C4 viene ora spenta anche la tensione del core e sono disattivate le memorie cache di primo e di secondo livello. In questo modo la Cpu è portata in uno stato a bassissimo consumo energetico, di fatto prossimo a 0 watt. Mentre si trova in questo stato il chipset di supporto al processore continua a gestire il traffico in memoria per le periferiche di I/O.
Bisogna considerare che il processo di ripristino di un core è tanto più oneroso quanto più è profondo il C-state il cui è posto. Transizioni troppo frequenti allo stato C4 e ritorno possono determinare anomali consumi e accorciare la vita delle batterie e quindi Intel ha previsto l'adozione di un algoritmo di tipo euristico che gestisce al meglio e solo quando è conveniente queste transizioni di stato.[38]

[37] www.pcprofessionale.it/stappro61/uploads/2008/06/201-art-intel.pdf
[38] http://www.intel.com/content/www/us/en/architecture-and-technology/microarchitecture/45nm-next-generation-core-microarchitecture-white-paper.html

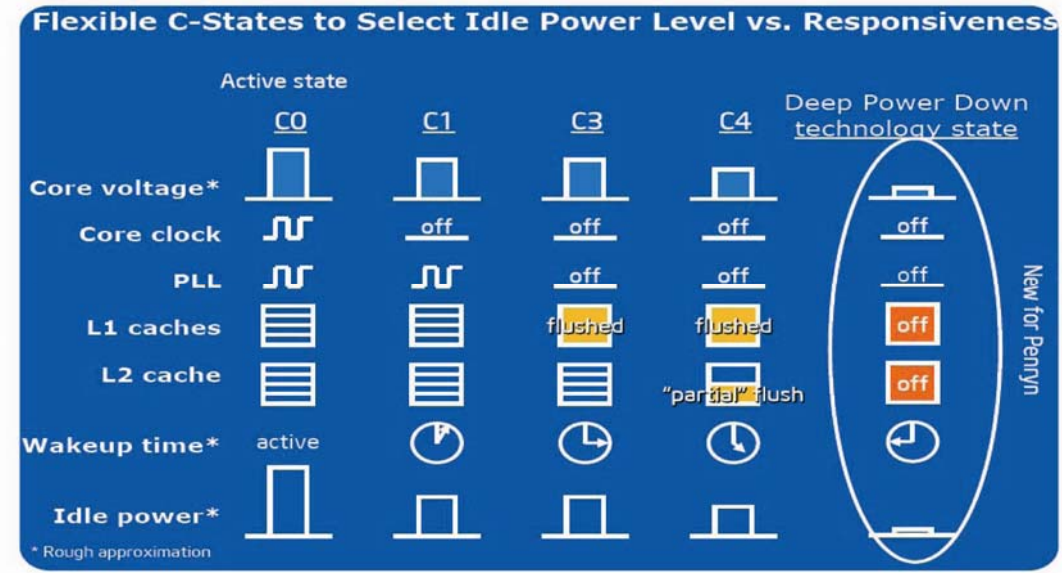

Figura 19 Gestione risparmio energetico su architettura Core e Penryn[39]

La seconda tecnologia introdotta da Intel in questo segmento prende il nome di **Enhanced Dynamic Acceleration Technology** (Edat) ed è una specie di overclock dinamico. Uno dei limiti dell'utilizzo dei processori multicore è che il software non sempre è ottimizzato per tali architetture. Per questo motivo è tutt'altro che raro avere un core completamente occupato e l'altro poco o per niente utilizzato (situazione tipica che si verifica durante l'utilizzo un'applicazione intensiva a singolo thread). Quando si verifica la situazione appena descritta, l'Edat innalza automaticamente la velocità di clock del core impegnato, incrementandone le prestazioni pur mantenendo il Tdp (Thermal Design Power) complessivo entro i limiti del processore.

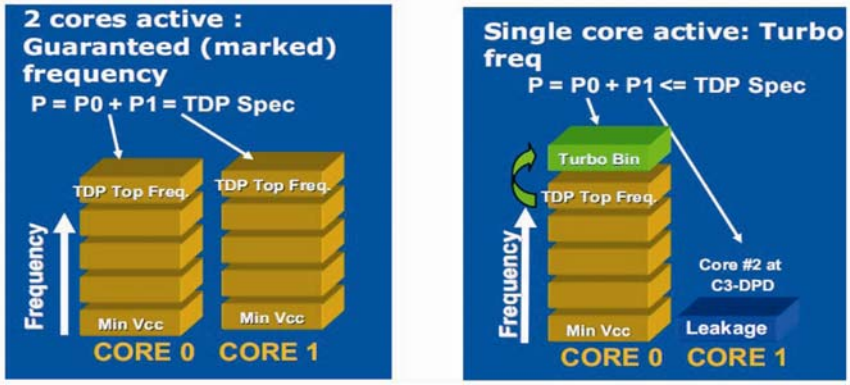

Figura 20 Overclock dinamico dei core nell'architettura Penryn

[39] download.intel.com/corporate/education/emea/event/af12/files/ronen.pdf

§ 1.2 Processori di nona generazione (2008 - Nehalem 45nm)

L'architettura Nehalem deriva direttamente dalla precedente Core e ne migliora molti aspetti soprattutto per quello che riguarda l'aumento delle prestazioni a parità di risorse energetiche utilizzate. Le novità apportate sono maggiormente focalizzate sull'ottimizzazione dei singoli core e sull'architettura complessiva per migliorare la sinergia operazionale tra i diversi core del processore a tutto vantaggio delle applicazioni, siano esse single threaded o multi threaded.[40]

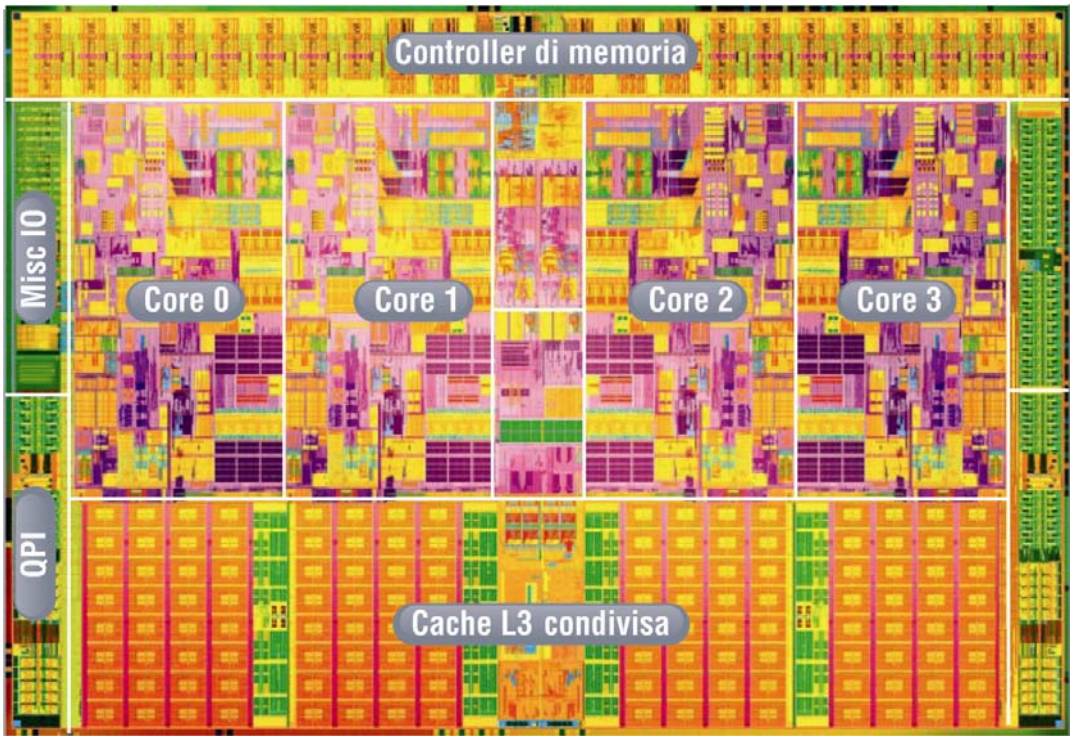

Figura 21 Si notano le novità introdotte da Intel: controller di memoria integrato, link Qpi, cache L3 e approccio quad core monolitico.[41]

Dalla figura si può vedere che per integrare i 4 core viene utilizzato un processo a **die monolitico** mentre in Penryn le versioni quad core sono realizzate con un'architettura di tipo Mcm (Multi Chip Module) in cui due blocchi dual core distinti sono montati in un unico package.

[40] Microarchitettura Nehalem
http://it.wikipedia.org/wiki/Nehalem_%28hardware%29
[41] Pc Professionale N. 212 (Novembre 2008) p. 102 - 109

Il vantaggio principale, di un approccio monolitico, è la possibilità di condividere alcune unità del processore (come la cache o altre unità di controllo) e di avere una connessione interna diretta tra i processori (con l'approccio a die doppio il collegamento tra i diversi die avviene tramite il bus di sistema). Lo svantaggio di realizzare un core monolitico è solo dal punto di vista costruttivo: mentre i quad core a doppio die sono realizzati accorpando due diversi die (e quindi possono essere testati separatamente e mixati per offrire le migliori performance) in un quad core monolitico questo non è possibile. Se uno dei quattro core ha dei problemi (frequenza di esercizio massima più bassa degli altri) o addirittura non è funzionante, il processore sarà scartato.

§ 1.2.1 Innovazioni della microarchitettura Nehalem

1 - QPI (**Quick Path Interconnect**)

Una nuova tipologia di Bus di sistema sostituisce il Fsb (Front Side Bus) che è il bus con cui sono trasferiti i dati tra la Cpu e il northbridge (chipset esterno al processore presente sulla scheda madre). Nelle architetture tradizionali, in cui il controller di memoria è integrato proprio nel northbridge, il Fsb è diventato un vero e proprio collo di bottiglia: cpu sempre più veloci o multicore necessitano di essere rifornite con sempre maggiore velocità di dati, per evitare di sprecare inutilmente cicli di clock per aspettare le informazioni. Lo scopo primario del bus QPI è quello di permettere al processore di comunicare direttamente con i vari altri componenti collegati alla motherboard, beneficiando quindi di una banda passante maggiore e di latenze sempre più ridotte. La sua caratteristica è quella di essere una tecnologia di connessione "point-to-point" che elimina gli svantaggi portati da un solo Bus condiviso tra tutti i processori il controller di memoria e il controller I/O.

Questa nuova architettura può trasferire dati fino a 25.6 GB/s. Ogni processore ha la sua memoria dedicata cui accede ad opera di un controller di memoria integrato. Nel caso in cui un processore necessiti di accedere alla memoria dedicata di un altro processore si utilizza il QPI ad alta velocità per il collegamento. Il bus QPI è di tipo bidirezionale con un canale da 20 bit in ciascuna direzione; di questi 20 bit, solo 16 sono riservati al trasferimento effettivo dei dati da elaborare, mentre gli altri 4

vengono utilizzati come bit di parità, ovvero come codici di correzione dell'errore, in maniera analoga a come avviene nelle memorie RAM con ECC.

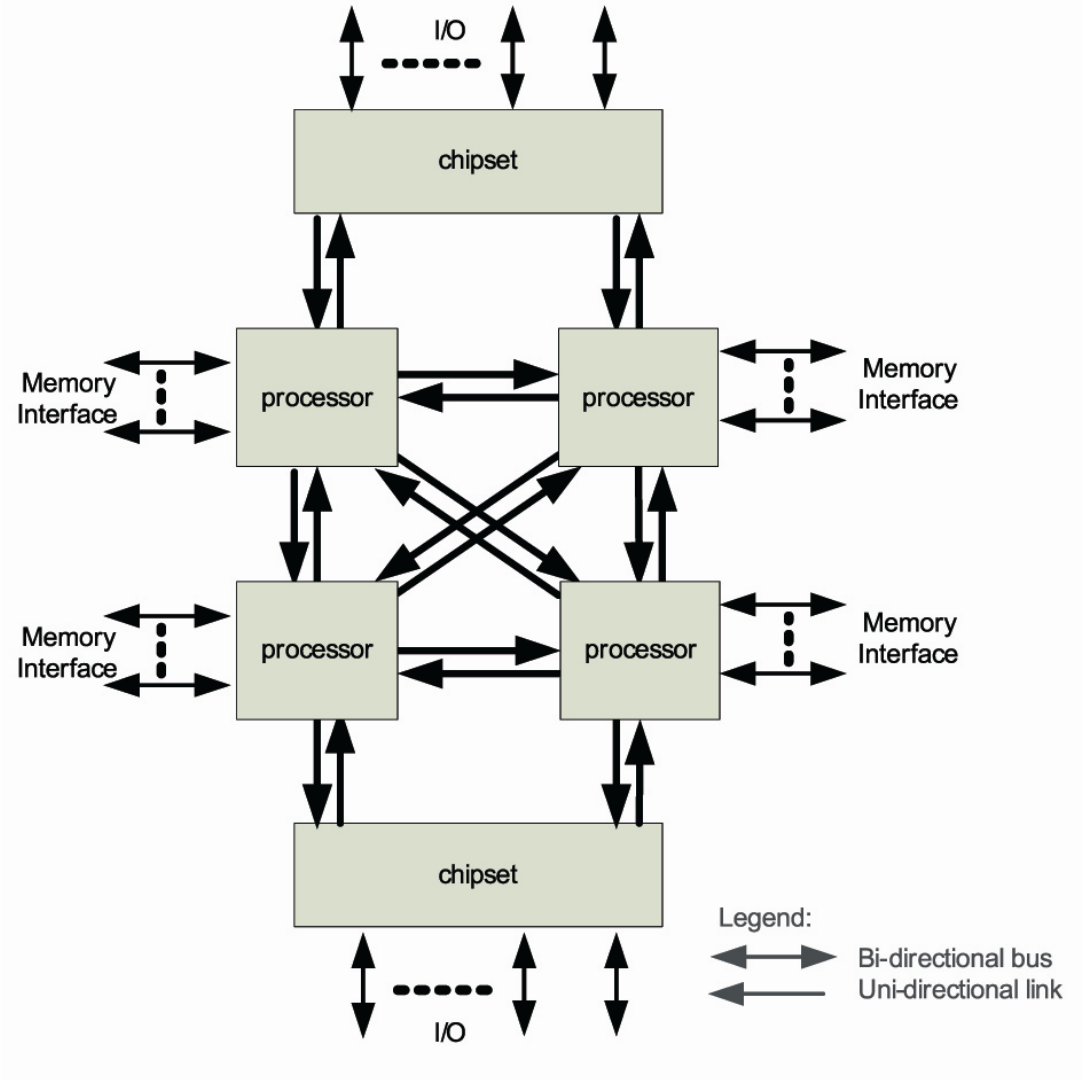

Figura 22 Intel Quick Path Interconnect [42]

Oltre che per la connessione tra Cpu e northbridge, il Qpi diventa anche il canale di connessione tra le diverse Cpu nelle piattaforme multiprocessore.

[42] An introduction to the Intel Quickpath Interconnect Jan 2009
www.intel.com/technology/quickpath/introduction.pdf

2 – Controller di memoria integrato nella Cpu

La presenza di un "memory controller" integrato nel processore consente di ottenere 2 importanti benefici: in primo luogo la riduzione della latenza relativa alle comunicazioni del processore con la memoria, in secondo luogo una scalabilità dell'ampiezza di banda direttamente proporzionale al numero di processori presenti nel sistema. Mediante l'approccio tradizionale, in cui il controller della memoria RAM era situato nel northbridge del chipset, ad ogni nuova CPU collocata nel sistema, l'ampiezza di banda del canale dedicato alla comunicazione con la memoria rimaneva costante, diventando anche un collo di bottiglia in certe configurazioni più spinte.

Con memorie DDR3-1333, il bandwidth disponibile in certe configurazioni diventa di 32 GB/s. Il vantaggio di avere un controller di memoria integrato non è solo una pura questione di bandwidth, perché abbassa sostanzialmente la latenza di accesso alla memoria, un aspetto importante se consideriamo che ogni accesso costa diverse centinaia di cicli. Nonostante la riduzione della latenza del controller di memoria integrato sia apprezzabile nel comparto desktop, saranno le configurazioni server multi-processore a beneficiare dell'architettura maggiormente scalabile. Mentre prima il bandwidth rimaneva costante quando venivano aggiunte le CPU, ora ogni nuova CPU aggiunta incrementerà il bandwidth, poiché ogni processore ha il proprio spazio locale di memoria.

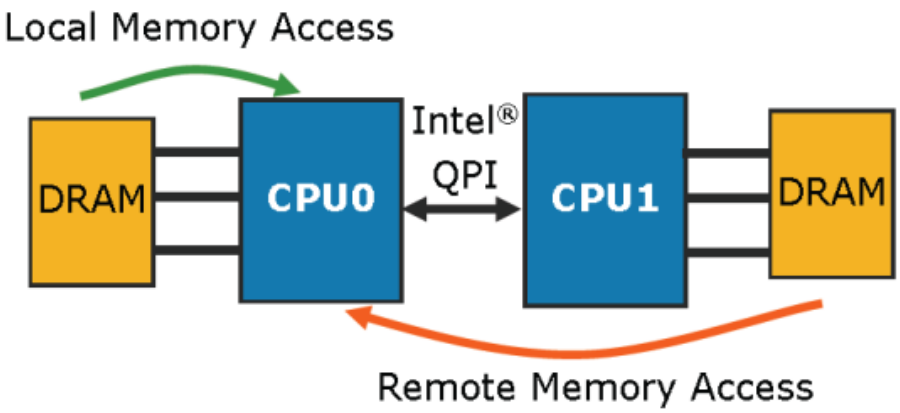

Figura 23 Modalità di accesso alla memoria alla memoria locale e a quella remota.

Questa è una configurazione Non Uniform Memory Access (NUMA), il che significa che gli accessi alla memoria possono essere più o meno importanti, a seconda dei dati contenuti nella memoria. Un accesso alla memoria locale ha ovviamente una latenza inferiore e un bandwidth maggiore; al contrario, un accesso alla memoria remota richiede il transito attraverso il collegamento QPI, che riduce le prestazioni.

3 – Intelligent Power Technology

Un aspetto per l'ottimizzazione energetica, è quello di minimizzare i consumi quando il processore è sotto-utilizzato. Mentre nelle generazioni di processori precedenti la riduzione di tensione passava per Vr (Voltage Regulator regolatori di tensioni) condivisi (e dunque per ridurre la tensione tutti i core dovevano essere in idle), adesso in Nehalem è previsto che ogni core disponga di Vr indipendenti. Questa soluzione prende il nome di **Integrated Power Gate** e prevede l'utilizzo di interruttori in grado di escludere completamente ogni singolo core dalla tensione di alimentazione. In questo modo ogni core può raggiungere fino allo stato C6 (quello che ha un consumo in pratica pari a 0 watt) indipendentemente dagli altri core, con bassi tempi di latenza. La supervisione di tutta la gestione del controllo energetico è demandata a un componente ben preciso, la **Pcu** (**Power Controller Unit**). Si tratta di un vero e proprio processore esterno alla cpu che la gestisce interamente dal punto di vista energetico fino al punto di arrivare ad essere l'unico componente attivo mentre la cpu è in stato di consumo minimo.

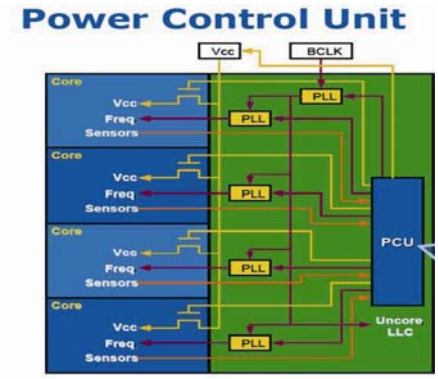

Figura 24 La Power Control Unit gestisce gli stati energetici dei core, fino allo spegnimento di ogni singolo core, e le frequenze di clock mediante l'utilizzo di sensori di temperatura,corrente e calore dissipato.

4 – Turbo Boost Technology

La funzionalità appena descritta può essere utilizzata per ottenere la massima performance del processore sfruttando lo spegnimento dei core non utilizzati e innalzando la velocità di clock di quelli in uso. Infatti uno dei limiti dell'utilizzo dei processori multicore è che il software non sempre è ottimizzato per tali architetture. Per questo motivo è tutt'altro che raro avere uno/due core completamente occupati e gli altri invece assolutamente inutilizzati (situazione tipica che si verifica durante l'utilizzo un'applicazione intensiva a singolo thread). Quando questo si verifica la Pcu provvede (in maniera istantanea e assolutamente trasparente per l'utente) fornire una tensione di alimentazione più alta al core occupato (in pratica fa un overclock) portandolo a frequenze di lavoro più elevate mantenendo il Tdp (Thermal Design Power) complessivo entro i limiti del processore.

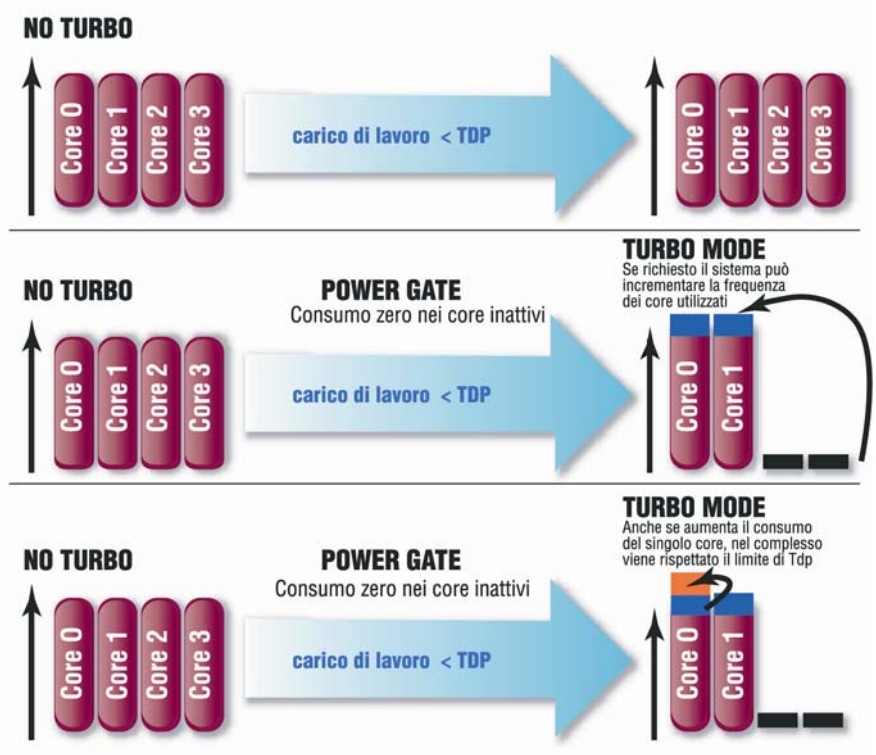

Figura 25 La tecnologia Turbo Boost permette di aumentare il clock dei core attivi pur rispettando i limiti imposti dal Tdp (thermal design power).

5 – Hyper-Threading Technology

Si tratta dell'evoluzione di terza generazione di una "vecchia" soluzione Intel, quella dell'Hyper-Threading (HT). In questo caso parliamo di **Simultaneous Multi-Threading** (SMT) che consente a ogni core fisico di elaborare contemporaneamente due distinti thread. Dal punto di vista logico un processore quad core è dunque equivalente a un'architettura a otto core. Per esmpio, mentre un thread è in attesa di un risultato o di un evento, un altro thread può essere eseguito dallo stesso core nella medesimo ciclo di clock riuscendo quindi ad utilizzare contemporaneamente quasi tutte le unità di calcolo presenti ed evitando sprechi di potenza elaborativa. Normalmente infatti i thread singoli non sfruttano tutte le unità di esecuzione, lasciando qualcuna di essa libera di essere usata, con HT attivo, da un secondo thread parallelo. In condizioni ideali si può guadagnare fino al 30% in più nelle prestazioni.

6 – Gestione migliorata della cache L3

La cache L1 rimane di 64 KB (divisa in due blocchi da 32 KB per le istruzioni e per i dati), mentre la cache L2, non condivisa e dedicata a ogni core, è di 256 KB. Sopra la L2 c'è un altro livello, la cache L3 di 8 MB, condivisa tra tutti i core. L3 ha un approccio di tipo inclusivo, ovvero racchiude le informazioni anche dei due livelli sottostanti (L2 e L1). In pratica nella cache L3 sono sempre occupati 1.280 Kbyte (64 x 4 di L1 più 256 x 4 di L2). Questo approccio, ha il vantaggio che ogni processore ha accesso alle informazioni delle cache di tutti gli altri processori direttamente nella L3 e, se vede che lì non sono disponibili, può direttamente passare alla memoria di sistema, senza dover scendere di livello e controllare anche le singole cache L1 e L2, con enormi benefici in termini di latenza.

§ 1.2.2 Evoluzione della Pipeline nella microarchitettura Nehalem

1 – 64 bit Macro Fusion

Rispetto all'architettura Core (v. rif. Pag 46) abbiamo un'evoluzione nel senso che vengono aggiunte nuove istruzioni che possono essere "fuse" insieme ed eseguite come una unica micro-istruzione (μop) e viene esteso il supporto alle istruzione a 64 bit.

2 – Enhanced Loop Stream Detector

Il "Loop Stream Detector" consente di avere in un buffer di dimensioni molto ridotte all'interno della Cpu tutte le istruzioni da elaborare all'interno del loop, evitando quindi l'accesso alla cache per ogni iterazione del loop. Con Nehalem Intel ha migliorato tale funzionalità portando da 18 a 28 le istruzioni che tale buffer è in grado di immagazzinare e inoltre esse non sono più di tipo x86 ma direttamente μops (micro-operations); un altro importante miglioramento si è avuto soprattutto spostando fisicamente all'interno della catena della pipeline la sua posizione; se nell'architettura Core esso era posizionato subito dopo la fase di "fetch" delle istruzioni (ovvero dopo la fase di lettura delle istruzioni da elaborare), in Nehalem si trova dopo la fase di "decode", in questo modo si possono disabilitare le unità di decodifica e di fetch in presenza di un loop, consentendo un maggiore risparmio energetico durante l'elaborazione di questi cicli.[43]

[43] Gabriel Torres "Inside Intel Nehalem Microarchitecture" (2006) p. 1-7
http://www.hardwaresecrets.com/article/Inside-Intel-Nehalem-Microarchitecture/535

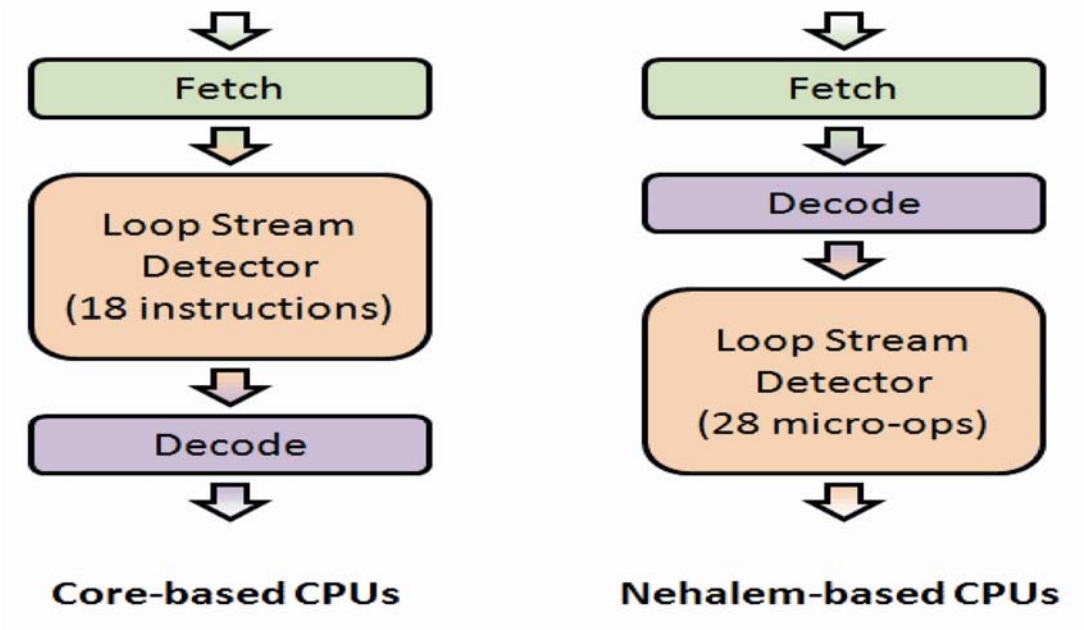

Figura 26 Localizzazione del Loop Stream Detector nelle architetture Core e Nehalem

2 – **Nuovo Translation Lookaside Buffer (TLB)**

Il **Translation Lookaside Buffer (TLB)** è un buffer, cioè una memoria tampone (in pratica una cache nella CPU), che l'MMU (Memory Management Unit) usa per velocizzare la traduzione degli indirizzi virtuali in indirizzi fisici.

L'utilizzo della memoria virtuale permette di allocare più memoria a un programma rispetto a quella fisicamente presente nel computer, posizionando parte dei dati necessari nella memoria fisica e il rimanente in un file (swap file) nell'hard disk.

La Memoria virtuale è segmentata in pagine di dimensioni prefissate e solo alcune pagine vengono caricate nella memoria fisica in zone dipendenti dalla politica di Page Replacement.

La Page Table (generalmente caricata in memoria) tiene traccia di dove le pagine virtuali sono caricate nella memoria fisica.

Il TLB è una cache della Page Table, cioè solamente un sottoinsieme del suo contenuto viene memorizzato. [44]

I processori della generazione precedente (Core2) usavano un TLB diviso in due parti, un TLB level 1 estremamente piccolo (16 entries) ma anche molto veloce per soli loads e un più grande TLB level 2 (256 entries) che amministrava e immagazzinava i loads missed nel TLB level 1.

Nehalem ha ora un vero TLB a due livelli: il TLB level 1 è condiviso tra dati e istruzioni. Il "TLB level 1 data" immagazzina 64 entries per small pages (4K) o 32 per large pages (2M/4M), mentre il "TLB level 1 instruction" immagazzina 128 entries per small pages (come il Core 2) e sette per large pages.

Il secondo livello "TLB level 2" è una cache unificata che può immagazzinare fino a 512 entries e opera solo con small pages. L'intento di questo miglioramento è incrementare le prestazioni delle applicazioni che usano grandi set di dati (database server ad esempio).

[44] Andrea Ferrario, "Intel Core i7 Nehalem : analisi architettura" (2008) http://www.tomshw.it/cpu.php?guide=20081016&page=architettura-nehalem-intel-corei7-09

§ 1.2.3 Evoluzione di Nehalem (2010 - Westmere 32nm)

Con Westmere, che di fatto è un'evoluzione di Nehalem e non una nuova microarchitettura, Intel introduce il processo produttivo a 32 nm utilizzando l'architettura di Nehalem.

La novità principale del nuova serie di microprocessori è l'integrazione all'interno della Cpu di un core grafico denominato Intel HD Graphics.

L'integrazione tra GPU e CPU, avviene a livello di package e non di silicio: sul package del processore, sono presenti due distinti die, uno per la componente CPU (die Westmere a 32nm) e l'altro per quella GPU, collegati tra di loro attraverso QPI, Quick Path Interconnect.[45]

Troviamo sullo stesso package due differenti die, con il più piccolo per la componente CPU costruito utilizzando tecnologia a 32 nanometri e il più grande che integra GPU e memory controller, costruiti invece con processo a 45 nanometri.

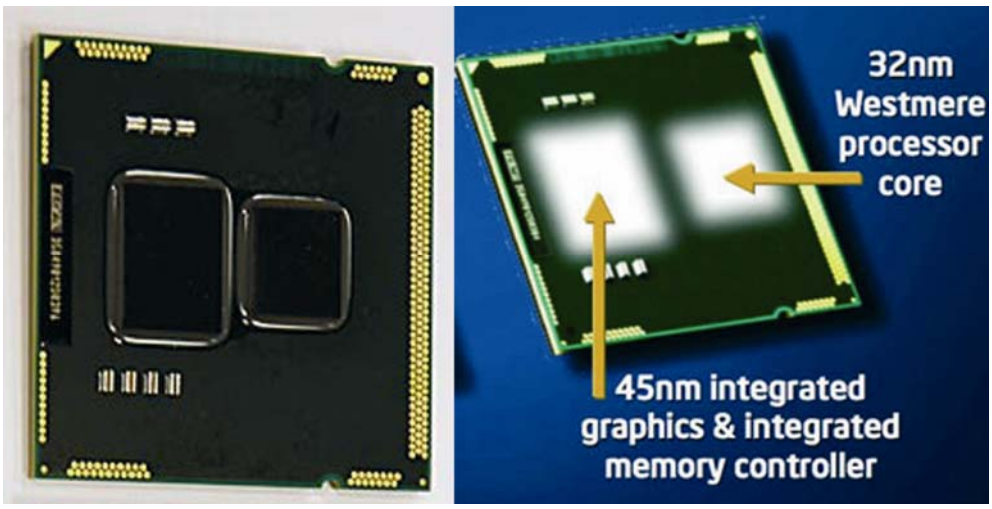

Figura 27 Processore della famiglia Westmere

L'integrazione del memory controller e della GPU sullo stesso package del processore ha permesso di semplificare l'architettura della motherboard riducendo il numero di componenti necessari sulla scheda madre. Da un approccio a 3 chip (CPU,

[45] http://www.hwupgrade.it/articoli/cpu/2353/intel-core-i5-661-le-prime-soluzioni-a-32-nanometri_2.html

north bridge e south bridge) si è passati ad uno a 2 chip, Cpu e chipset di supporto denominato **Pch** (Platform Controller Hub) che significa centro di controllo della piattaforma. La Cpu e il Pch comunicano tra loro per mezzo di due canali: quello **Dmi** (Direct Media Interface) assolve allo scambio dati generale fornendo una banda di trasmissione massima teorica pari a 10 Gbit al secondo in modalità full duplex; il nuovo **Fdi** (Flexible Display Interface) assolve invece al trasferimento delle informazioni video provenienti dal core HD Graphics alla logica di controllo che pilota i segnali sulle uscite Vga, Dvi, Hdmi o DisplayPort a seconda della situazione richiesta.

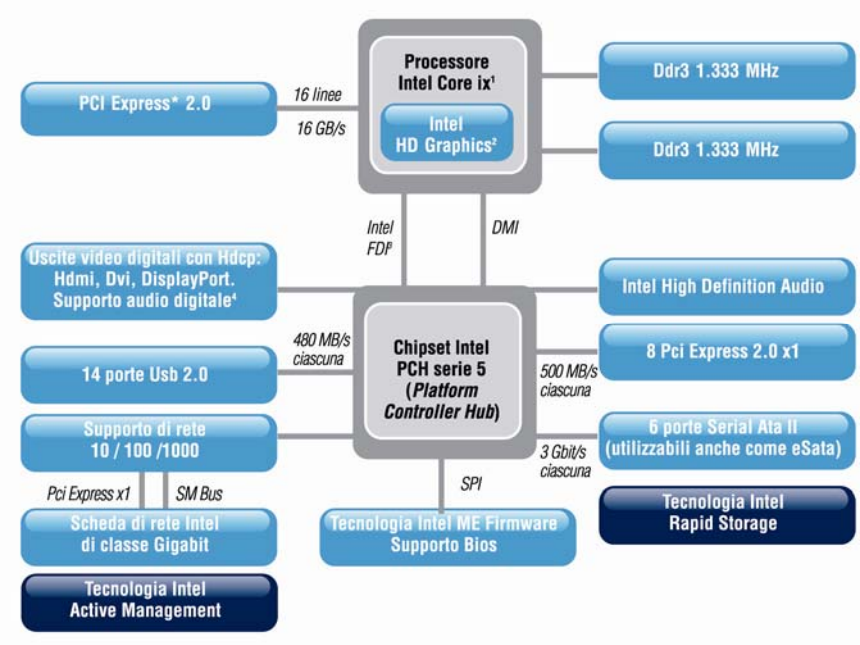

Figura 28 Schema del nuovo chipset della serie 5 per processori della famiglia Westmere[46]

Le caratteristiche base dell'evoluzione Westmere indicano che per ogni core sono presenti una cache L1 da 64 KB e L2 da 256 KB, affiancate da una cache L3 di 3MB o 4 MB condivisa tra i vari core.

Con Westmere debuttano 6 nuove istruzioni, denominate AES-NI (AES New Instructions), che forniscono l'accelerazione in hardware per alcune funzioni di protezione basate sulla tecnologia Aes (Advanced Encryption Standard). Il supporto hardware fornito dalle istruzioni AES-NI permette di ridurre il carico di lavoro sul sistema e di recuperare parte della potenza di calcolo del processore. Oltre alla

[46] Pc Professionale N. 227 (Febbraio 2010) p. 24 - 35

codifica di file e dischi rigidi, l'utilizzo della tecnologia Aes è impiegata per la messa in sicurezza del traffico di rete e può risultare utile per cifrare comunicazioni di tipo Voip.

La presenza delle nuove istruzioni AES-NI permette anche di proteggere la Cpu dalle vulnerabilità tipiche di un algoritmo che fa uso di tabelle di look up per la codifica. Data la natura dipendente dai dati della tabella di look up un programma che analizza i tempi di accesso della cache può rapidamente trovare le chiavi di codifica. Dal momento che la latenza di accesso delle istruzioni AES-NI, essendo fissa, non dipende dal dato, questo tipo di attacco diventa inutile.[47]

L'implementazione della tecnologia Turbo Boost anche per il comparto grafico assicura una migliore gestione della potenza elaborativa dell'intera piattaforma. Nel momento in cui un'applicazione necessita più potenza da parte della Gpu rispetto a quella della Cpu, la frequenza operativa del chip grafico integrato può essere innalzata fino a raggiungere il limite di consumo massimo previsto dal parametro Tdp. Quando invece il comparto grafico si trova a svolgere la sola visualizzazione del desktop o di applicativi 2D la frequenza può essere ridotta lasciando margine per l'incremento di quella dei core all'interno del die Westmere.

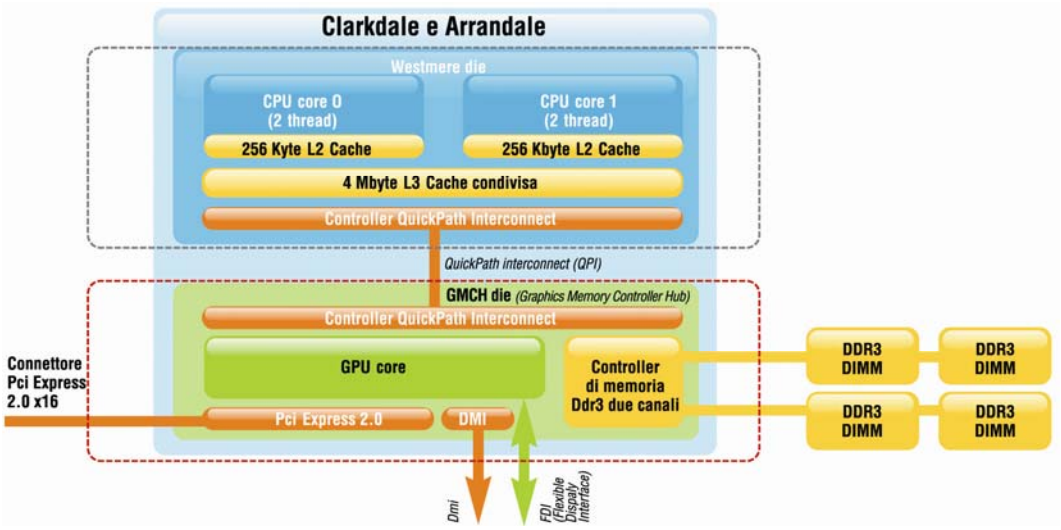

Figura 29 Struttura di un processore dual core della famiglia Westmere. Il die Westmere assolve alle funzioni di Cpu vera e propria, mentre quello HD Graphics contiene al suo interno il motore grafico, il controller di memoria e il sistema di gestione del bus Pci Express 2.0 primario che può essere impiegato per collegare una scheda grafica di tipo discreto.

[47] D. Kanter : Westmere arrives (2010)
http://realworldtech.com/page.cfm?ArticleID=RWT031710140138&p=1

§ 2 – Microarchitetture attuali

§ 2.1 Processori di decima generazione (2010 - Sandy Bridge 32nm)

Sandy Bridge è la nuova microarchitettura realizzata con l'ultima tecnologia consolidata (quella a 32nm utilizzata per la produzione di Westmere). Si tratta quindi della microarchitettura corrispondente alla nuova fase "tock" nel modello di sviluppo dei microprocessori ideato da Intel.

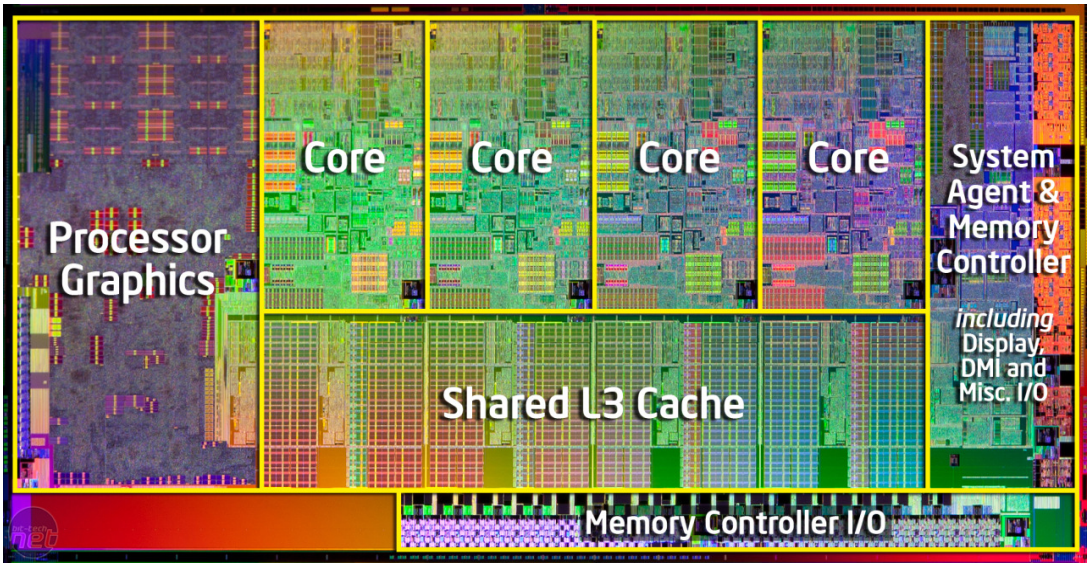

Figura 30 Il die di Sandy Bridge include un processore grafico Igp e un blocco di gestione (System Agent) centralizzato. La cache L3 chiamata ora LLC (last level cache) è condivisa tra i core presenti e il processore grafico. L'accesso alla cache avviene tramite il nuovo bus interno denominato Ring Bus.[48]

L'integrazione della GPU ha portato ad un aumento nel numero di transistor, che nella declinazione quad core in figura sfiora il numero di 1 miliardo.[49] Per ogni core integrato nei processori Sandy Bridge Intel ha utilizzato circa 55 milioni di transistor, che si vanno ad abbinare ai 114 milioni riservati alla GPU integrata. La tecnologia produttiva è a 32 nanometri, con transistor Hi-K metal gate di seconda generazione, mentre la cache LLC (ex L3) può giungere sino a un quantitativo massimo di 8 Mbytes, per la prima volta unificata tra CPU e GPU.

[48] Pc Professionale N. 239 (Febbraio 2011) p. 26 - 39

[49] http://www.hwupgrade.it/articoli/cpu/2674/intel-sandy-bridge-analisi-dell-architettura_index.html

2.1.1 - Evoluzione dei Core (Front End e Back End)

Le soluzioni desktop prevedono modelli Core i7 (4 core), Core i5 (4 core) e Core i3 (due core). Gli elementi discriminanti tra le varie proposte sono il supporto alla tecnologia Turbo Boost, assente nei modelli Core i3, il numero massimo di threads che possono venir processati, sempre pari a 8 nelle soluzioni Core i7 e a 4 in quelle restanti, la dimensione della cache L3 unificata, la frequenza di clock e il TDP massimo che non si spinge oltre i 95 Watt.

Ciascun core è composto da un Front End che si occupa di fornire dati e istruzioni al Back End in cui risiede l'Execution Cluster deputato all'esecuzione vera e propria delle istruzioni.

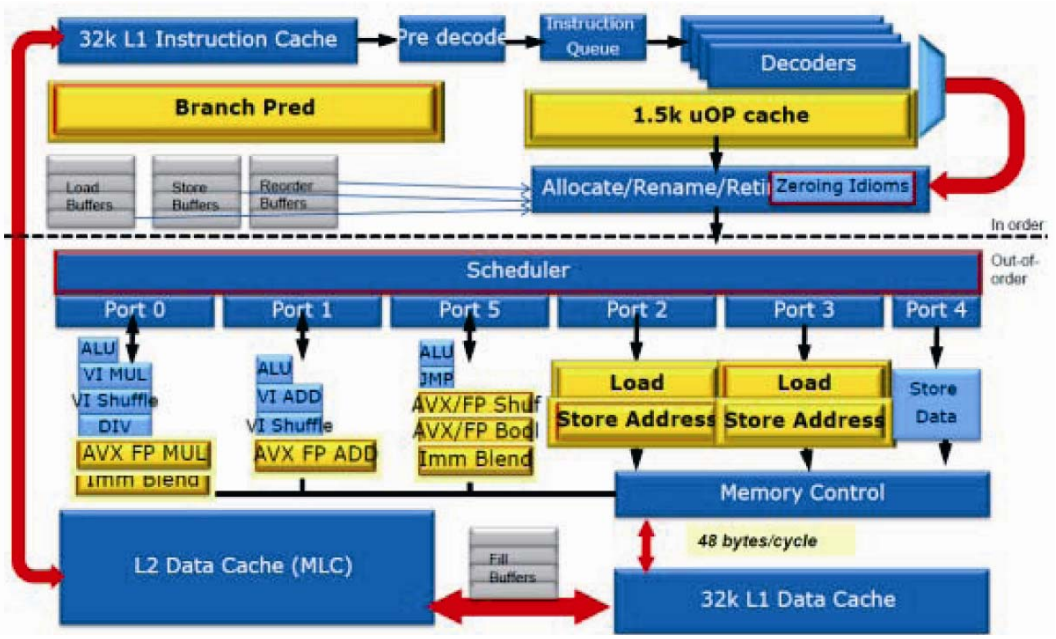

Figura 31 Front End e Execution cluster

A partire dalle CPU Core 2 Intel ha introdotto il Loop Stream Detector o LSD, componente che verifica che il processore stia eseguendo un loop software; quando questo avviene, il branch predictor e le unità di fetch e decode possono venir spenti così da risparmiare energia, inviando dati alle execution units attraverso cosiddette micro-ops che sono state inserite nella cache direttamente dal Loop Stream Detector. Sandy Bridge migliora questo approccio grazie ad una cache specifica per le micro-op, indicata come Decoded Uop Cache.

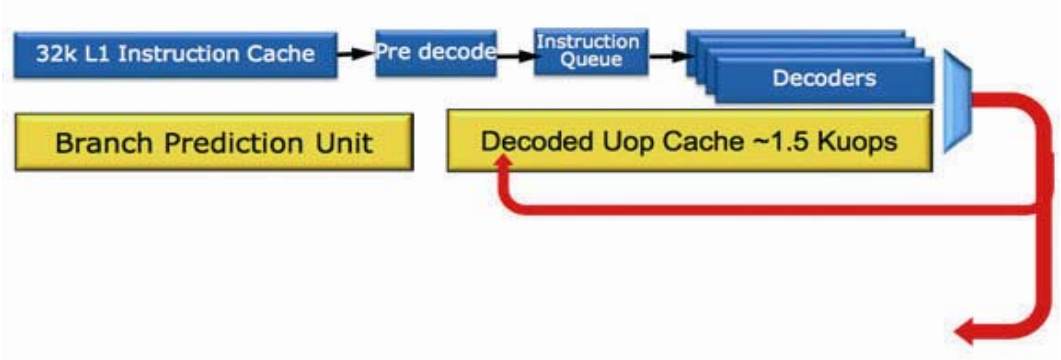

Figura 32 Posizione della Decoded Uop Cache nel front end

La Decoded Uop Cache prende le istruzioni dopo che sono state decodificate inserendole in una sorta di cache L0: nel momento in cui una istruzione è presente è la stessa cache a inviare i dati nella pipeline, permettendo al processore di spegnere del tutto il front end con un diretto beneficio in termini di consumo complessivo. La cache è di 6KB e può memorizzare al massimo circa 1.500 micro-ops, ottenendo secondo Intel indicativamente l'80% di hit rate per la maggior parte delle applicazioni. Un'altra caratteristica è il fornire alla pipeline i dati in modalità sostenuta, con un impatto positivo sulla latenza.

Altra novità è rappresentata dalla **nuova unità di branch prediction** rivista completamente. Se infatti la branch prediction unit è in grado di prevedere senza errori (o meglio, con pochi errori) i salti, la CPU evita di svuotare continuamente la pipeline per gestire il cambio di contesto e perciò può garantire prestazioni elevate.
Per rendere tale unità più efficiente Intel l´ha completamente riprogettata utilizzando buffer per la memorizzazione degli indirizzi di salto e della storia delle previsioni maggiormente densi di dati. Ciò permette di tenere in memoria una storia dei salti più lunga senza incrementare la dimensione della struttura dati e dunque occupare più spazio/consumare di più. Il produttore afferma che con queste modifiche la correttezza dell´unità di branch prediction è stata migliorata del 5% rispetto alle CPU Core di prima generazione.[50]
Le modifiche più pesanti sono state però apportate al cluster Out-of-Order.

[50] Dino Fratelli, Intel Sandy Bridge: i segreti dell´architettura Core 2nd Gen (Gennaio 2011) http://www.dinoxpc.com/articolo/Intel+Sandy+Bridge%253A+i+segreti+dell%2526acute%253Barchitettura+Core+2nd+Gen_1170.htm

In Sandy Bridge, al posto del Retirement Register File introdotto con l´architettura Nehalem è stato utilizzato un **Physical Register File**: il primo prevede che i dati necessari alle micro-ops, riorganizzate secondo il funzionamento OOO, vengano copiati nei registri per ogni operazione effettuata. Con la nuova microarchitettura, invece, viene salvato solo un link (puntatore) ai dati presenti nel registro fisico. Il registro fisico è inoltre suddiviso in un blocco per la gestione dei valori integer (160 voci) e in un blocco per i valori floating point (144 voci). Queste scelte permettono di evitare la duplicazione di dati e ridurre il traffico degli stessi con vantaggi sia in termini di performance che di consumi. Lo svantaggio è nell´introduzione di un ulteriore stadio nella execution pipeline che serve a dereferenziare il puntatore.

L´utilizzo di un Physical Register File è stato necessario anche per gestire in maniera efficiente le nuove istruzioni AVX a 256-bit: trasferire un così importante quantitativo di dati in ingresso e uscita avrebbe comportato un serio decadimento delle prestazioni. Le istruzioni AVX (Advanced Vector Extensions), come già detto, rappresentano un´ulteriore estensione del set SSE, e utilizzano registri vettoriali SIMD (singol instruction multiple data) ampi 256-bit e prevedono l´esecuzione non distruttiva delle operazioni (ad esempio, il risultato di a=a+b non viene memorizzato nel registro a ma in un terzo registro c), il tutto a garanzia di una semplificazione di molti algoritmi complessi per migliorare il rapporto performance per watt.

Rispetto a Nehalem, il buffer delle istruzioni riordinate (ROB) è in grado di memorizzare fino a 168 micro-ops contro le precedenti 128, mentre il numero di Load e Store buffer cresce da 48 e 32 a 64 e 36 rispettivamente.

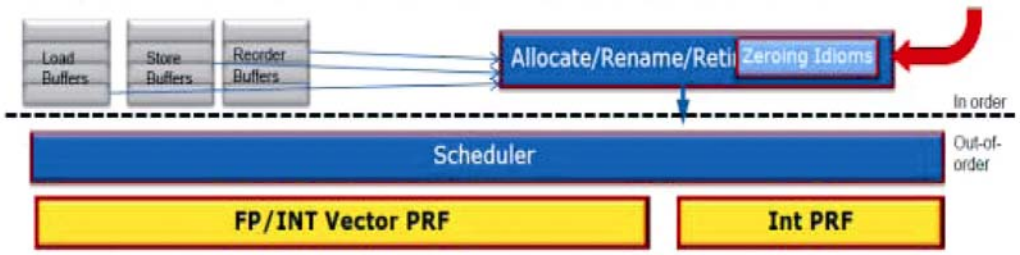

Figura 33 Physical Register File (PRF) al posto del Retirement Register File introdotto in Nehalem

Per abilitare l'esecuzione delle nuove istruzioni AVX, Intel ha dovuto lavorare anche sul fronte **Execution Cluster** in modo da supportare operandi a 256 bit senza dover raddoppiare in modo completo l'architettura precedente.

Le tre execution port di Nehalem sono state conservate, aggiungendo la possibilità di usarle in maniera diversa da quella tradizionale. Ogni porta dispone infatti di unità per lavorare simultaneamente con tre tipi di dato, Integer a 64-bit, Integer a 128-bit e Floating Point a 128-bit: per gestire un´istruzione AVX a 256-bit sarebbe dunque possibile utilizzare due path a 128-bit, uno integer e l´altro floating point.
Quando viene richiesta l'esecuzione di una istruzione AVX due data Path vengono utilizzati simultaneamente in modo da formarne uno in grado di supportare gli operandi a 256 bit. Grazie a questo espediente è stato quindi possibile un risparmio ingente in termini di transistor, risparmio che è stato investito in un incremento dei registri (da 128 a 168) a supporto del riordino delle istruzioni.

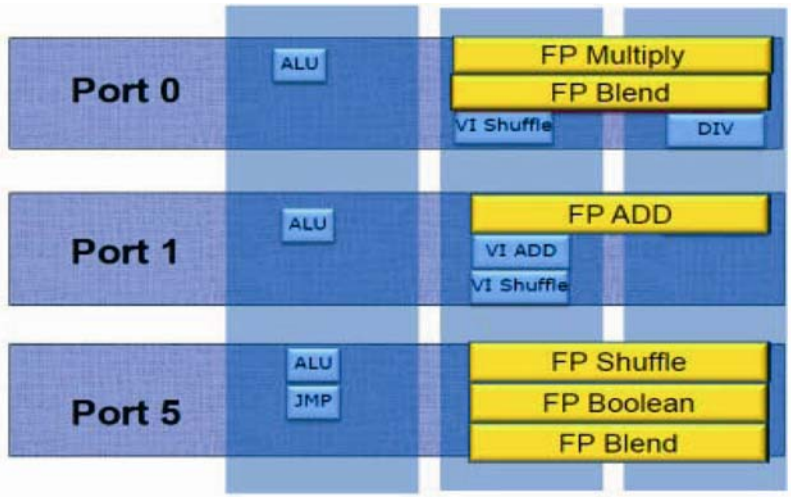

Figura 34 Execution cluster e riorganizzazione dei data path

Intel ha modificato anche il **Memory Cluster** per quanto riguarda le unità di Load e Store della CPU che ora prendono in carico un maggior quantitativo di dati: Sandy Bridge utilizza due accessi unificati che permettono di gestire operazioni di caricamento (dati e indirizzi) e memorizzazione (indirizzi). Il terzo accesso non ha subito cambiamenti rispetto a quanto s'incontra nell´architettura Core di prima generazione ed è utile per la memorizzazione dei dati. Ognuno dei suddetti path è in grado di veicolare fino a 16 byte (128-bit) di dati per ciclo di clock garantendo così un throughput complessivo della cache L1 dati superiore del 50% rispetto a Nehalem: Sandy Bridge può gestire due richieste di lettura per un totale di 32 byte di dati ed una in scrittura per un totale di 48 byte di dati per ogni ciclo di clock.

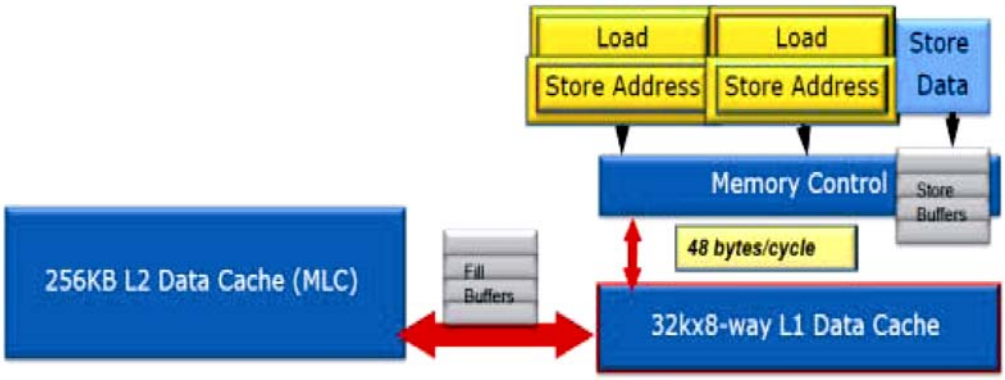

Figura 35 Riorganizzazione dei datapath nel Memory Cluster

2.1.2 - Evoluzione del Core Grafico

Una notevole evoluzione in termini di prestazioni deriva dall'adozione di un processore grafico incorporato Igp (Integrated Graphics Processor) denominato Intel Graphics Media Accelerator HD. Data la sua profonda integrazione con il silicio della Cpu, l'Igp può essere considerato come un core specializzato nello svolgimento di compiti dedicati alla grafica e all'elaborazione video.

Rispetto alla microarchitettura precedente (Clarkdale e Arrandeale) in cui il processore grafico era inserito nello stesso package del processore ma in un die distinto con tecnologia a 45nm rispetto a quella a 32nm della Cpu e quindi con benefici limitati, la profonda integrazione di Cpu e Igp di Sandy Bridge non subisce le limitazioni di prestazioni dovute al diverso processo produttivo e alla separazione dal die della Cpu.

La particolarità di questa soluzione è l'inserimento nel silicio di istruzioni di codifica e decodifica del flusso video tali che il nuovo Igp potrebbe meglio essere definito come un core video piuttosto che un core grafico. L'adozione di una soluzione "ibrida" composta da unità di elaborazione programmabili e blocchi di esecuzione a funzioni fisse ha permesso di dar vita a un Igp con prestazioni molto superiori a quelle della precedente generazione arrivando a gestire la grafica 3D ad un livello paragonabile a quelle di una scheda grafica discreta di buon livello presente nel sistema.

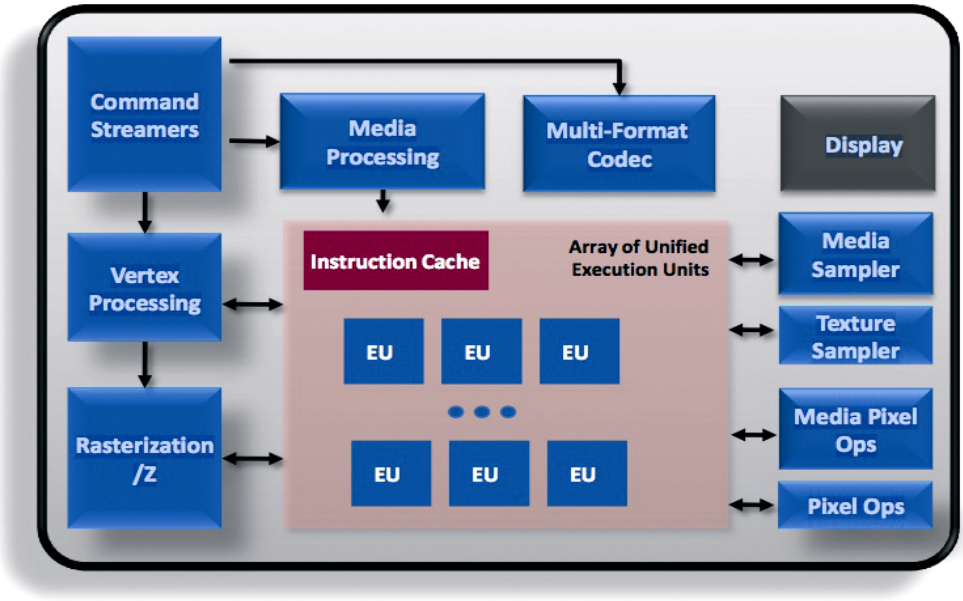

Figura 36 All'interno del Graphics Media Accelerator HD è presente un logica programmabile le cui unità di elaborazione Eu (Execution Unit) sono 6 per la versione HD2000 e 12 per la versione HD3000. A fianco trovano posto blocchi operativi a funzioni fisse molto efficienti e veloci nell'esecuzione ma non programmabili.

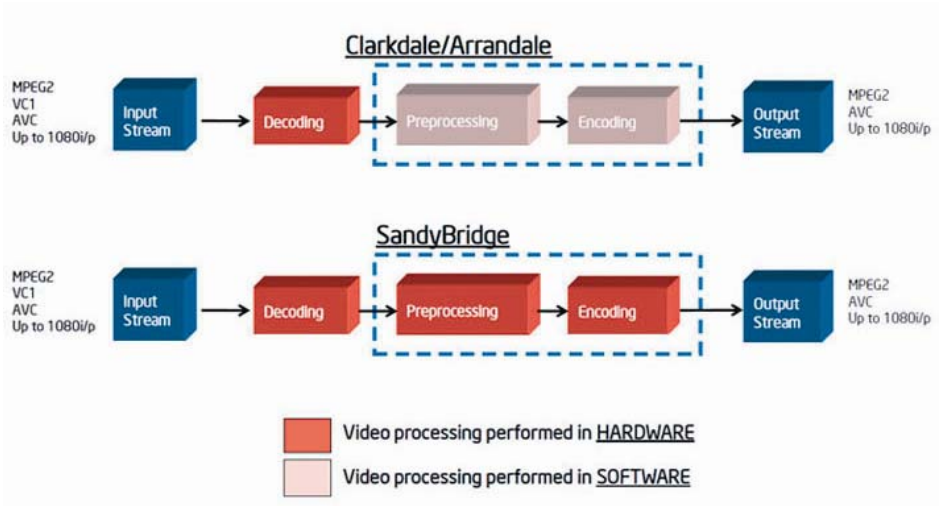

Figura 37 Media processing. Se confrontiamo le due soluzioni grafiche offerte da Clarkdale e Sandy Bridge risulta evidente come quest'ultima, avendo l'integrazione in hardware dei blocchi di elaborazione e codifica, permetta di ottenere un notevole incremento di prestazioni soprattutto nelle operazioni di transcodifica di video ed altri contenuti multimediali.

Bisogna sottolineare che gli ottimi risultati che si ottengono nel settore del video processing sono dovuti anche alla capacità dell'Igp di comunicare in modo diretto con i core di calcolo, con la velocissima Llc condivisa con i core e con il controller di memoria integrato nella Cpu.

La frequenza di funzionamento dell'Igp è variabile dal momento che esso opera su un piano energetico indipendente da quello dei core di calcolo. Ciò permette di applicare la tecnologia Turbo Boost a questo componente con una frequenza di lavoro che, a seconda del processore, può variare da 350 Mhz a 1.300 Mhz.

2.1.3 - System Agent

All'interno del System Agent (v. fig.32) trova posto il controller di memoria DDR3 a doppio canale (bus a 128bit velocità fino a 2.133 Mhz) al quale hanno accesso sia i core che la grafica integrata, l'interfaccia Dmi (Direct Media Interface) che collega il processore al chipset esterno, il controller Pci Express 2.0 a 16 linee, l'unità di gestione del risparmio energetico Pcu (Power Control Unit) e il Display Engine per il trasferimento alle uscite video integrate sulla scheda madre del segnale video della grafica integrata.

2.1.4 – LLC Last Level Cache

La cache di terzo livello permette di incrementare le prestazioni mantenendo all'interno della Cpu un maggior numero di dati e istruzioni. La cache L3 di Sandy Bridge è stata rinominata Llc (Last Level Cache) ed è accessibile sia ai core classici che al core grafico. La sua struttura non è più a blocco unico ma è ripartita in tanti banchi quanti sono i core presenti. Il quantitativo di memoria Llc presente varia in base al modello di processore :

- 8MB (ripartiti i 4 blocchi da 2MB) per le unità Core i7
- 6MB (ripartiti in 4 blocchi da 1,5MB) per le unità Core i5
- 3MB (ripartiti in 2 blocchi da 1,5MB) per le unità Core i3.

La suddivisione della cache L3 in banchi permette di aumentarne la banda dati (i banchi lavorano in parallelo usando ognuno una connessione per raggiungere una banda dati teorica massima di 384GB/s nelle CPU quad-core e 192GB/s in quelle

dual-core) ma soprattutto garantire che essa scali al crescere del numero di core (e dunque di banchi).

Dal momento che la cache Llc è integrata nel die insieme ai core di calcolo, la sua frequenza operativa varia in sintonia con quella dei core stessi. Al crescere della frequenza dei core, attraverso l'uso della tecnologia Turbo Boost, la velocità della cache Llc cresce e permette di sostenere il maggior numero di operazioni che i core di calcolo sono in grado di eseguire; quando i core entrano in fase di idle la frequenza della cache scende (se non ci sono richieste da parte del core grafico) riducendo i consumi complessivi del processore.

2.1.5 – Architettura Ring Bus

Il sistema di bus interni necessari alla comunicazione dei vari blocchi funzionali interni ad una Cpu Sandy Bridge cambia radicalmente rispetto alle microarchitetture precedenti con l'introduzione del **Ring Bus** (bus ad anello) che garantisce una elevata scalabilità della struttura di trasferimento dati al crescere degli agenti collegati ad essa. Con il termine agenti si identificano tutti i blocchi logici che hanno accesso al bus di comunicazione e quindi i core, il processore grafico, il System Agent, la cache Llc.
Il ring è di tipo **fully pipelined**, pertanto opera alle stesse frequenze di clock e tensione della componente core.
Il ring non è bidirezionale, ma i punti di stop presenti all'altezza dei core, della GPU e del System Agent permettono di fatto di andare in direzione up e down. Il ring bus gestisce coherency, ordering e core interface per bilanciare al meglio la distribuzione delle richieste all'interno del ring.
Tutte le unità funzionali sono dunque connesse mediante un bus ad anello che utilizza un protocollo di comunicazione simile a quello del bus QPI (bus seriale punto-punto): in questo modo viene minimizzato il numero di connessioni interne al processore per il routing dei segnali e garantita un´elevata banda passante, pari a 96GB/s per ogni connessione.

Quando un agente vuole “comunicare” con un altro agente pone l’informazione sull’anello e questa informazione viaggerà sull’anello fino a raggiungere la sua destinazione. Il ring bus è in effetti costituito da 4 anelli a 256bit che sono identificati come Snoop, Request, Aknowledge e Data.

Ciascun agente accede al Ring Bus attraverso una stazione e ciascuna stazione tocca l’anello sia nella sua fase ascendente che nella sua fase discendente. In questo modo il Ring Bus mette a disposizione sempre il percorso più breve per le richieste e le informazioni permettendo di saltare da un punto all’altro dell’anello durante il transito attraverso una delle stazioni.

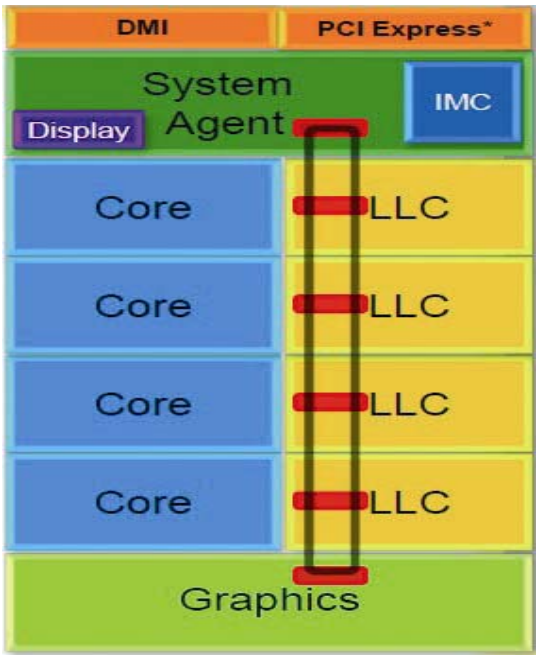

Figura 38 Il Ring Bus collega tutte le unità funzionali della cpu Sandy Bridge

Gli accessi al ring sono configurati in modo tale da seguire sempre il path più corto, minimizzando la latenza complessiva. Intel dichiara una latenza della cache LLC variabile da da 26 a 31 cicli di clock, inferiore ai 36 cicli di clock di media delle soluzioni Nehalem.

Il ring bus fa sì che anche il core grafico sia connesso non direttamente al controller delle memorie ma alla cache Llc. I vantaggi sono evidenti: performance più elevate che derivano dal fatto che il questo componente non va ad occupare parte del bus della memoria di sistema sottraendolo ai bisogni dei core della CPU e dal fatto che lo

scambio di dati fra Cpu e Igp può avvenire all´interno del processore stesso usando la cache Llc come memoria di scambio.

2.1.6 – Tecnologia Turbo Boost 2.0

Il System Agent contiene anche il blocco PCU - Power Control Unit - unità programmabile che raccoglie tutte le informazioni circa lo stato fisico di diverse sezioni della CPU come temperature e correnti per controllare di conseguenza frequenze e tensioni.

All´interno del blocco PCU sono implementate le funzionalità del risparmio energetico e della modalità Turbo Boost 2.0.

Intel ha suddiviso Sandy Bridge, da questo punto di vista, in tre blocchi con altrettanti algoritmi di gestione del clock e dell´alimentazione indipendenti.

Il primo contiene processore e cache L3 che funzionano, perciò, agli stessi valori di frequenza e tensione.

Il secondo include il solo core grafico che funziona alla sua specifica frequenza.

Il terzo è invece rappresentato dal System Agent stesso.

La tecnologia Turbo Boost parte dalla considerazione che le specifiche operative di un processore sono calcolate in modo tale che quando questo opera al massimo delle sue possibilità il suo consumo e le sue temperature di esercizio raggiungano un limite prestabilito identificato dal parametro TdP (Thermal Power Design). Dal momento che il processore si trova raramente a operare in questo stato limite, il margine energetico residuo viene impiegato dalla tecnologia Turbo Boost in modo intelligente per innalzare la frequenza operativa dei core effettivamente attivi per fornire così un maggior livello di prestazione degli stessi.

Con la versione 2.0 Intel modifica ulteriormente il limite di questa tecnologia basandosi sullo studio della velocità del transitorio che porta al punto di saturazione per il TdP. Nel momento in cui un processore passa dallo stato di idle a quello di

utilizzo a pieno carico si genera un incremento della temperatura che non è istantaneo sino al valore massimo, ma avviene gradualmente. Questo implica che per un certo lasso di tempo (calcolato al massimo in 25 secondi) la CPU abbia un margine di incremento della frequenza di clock, via tecnologia Turbo Boost, che è superiore rispetto a quanto disponibile nel momento in cui il processore è da un certo lasso di tempo in una condizione di pieno carico. La CPU, quindi, è portata ad una frequenza di clock ancora più alta di quella predefinita dalla tecnologia Turbo Boost per un periodo di tempo ridotto ma tale in ogni caso da velocizzare la risposta del sistema per applicazioni intensive che non si vanno a protrarre a lungo nel tempo. Il System Agent monitora in tempo reale le tensioni di alimentazione, le correnti assorbite e le temperature di tutte le zone del die e applica l'aggiustamento delle frequenze operative ai vari core in modo tale che al raggiungimento della soglia termica definita dal Tdp la frequenza turbo applicata sia scesa a quella di regime dopo il picco applicato in fase iniziale.

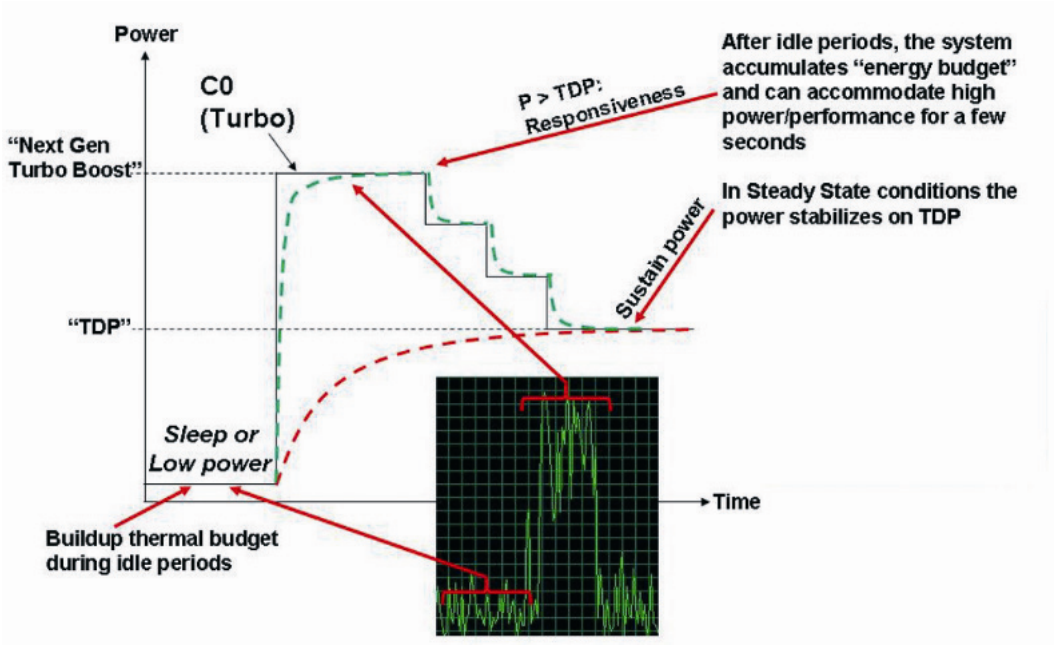

Figura 39 Tecnologia Turbo Boost 2.0. Rispetto al Tdp nominale che può essere protratto nel tempo, è possibile applicare una frequenza di clock iniziale più elevata per un periodo non superiore a 25 secondi per scendere poi a frequenze minori rientrando nei limiti del Tdp.

§ 3 – Microarchitetture future

3.1 – Evoluzione di Sandy Bridge (2012 – Ivy Bridge)

Ivy Bridge rappresenta l'evoluzione al processo a 22nm dell'architettura Sandy Bridge e non sarà solo un die-shrink (ridimensionamento del die della cpu) ma segnerà l'introduzione del nuovo transistor tridimensionale tri-gate 3D, che rappresenterà il futuro dell'intera linea di microprocessori, in grado di offrire una combinazione senza precedenti di risparmio energetico e miglioramenti prestazionali (v. pag.40).

Ivy Bridge sarà proposto ancora in **configurazioni a due e quattro core** (con grafica integrata) e per diversi modelli sarà presente l'Hyper-Threading. Supporterà il **PCI Express 3.0** anche se è improbabile un incremento del numero di linee.[51]

Sul fronte grafico avremo la presenza di un **core integrato con supporto DirectX 11 (Pc) e OpenCL 1.1 (Mac)**, dotato di 16 unità di esecuzione rispetto alle 12 presenti nella HD Graphics 3000, e saranno migliorate ulterioriormente le operazioni di codifica e decodifica video. La GPU potrà inoltre gestire fino a tre display e garantire prestazioni superiori nell'esecuzione dei contenuti multimediali grazie alla nuova versione di **Quick Sync Video** che raddoppia le prestazioni rispetto alla versione in Sandy Bridge**.**

Nonostante il miglioramento delle prestazioni grafiche verrà ridotto il clock della GPU e utilizzando un voltaggio inferiore, grazie all'utilizzo del processo di fabbricazione a 22nm con tecnologia 3D tri-gate, le prestazioni per watt della GPU Ivy Bridge raddoppieranno rispetto a quelle della GPU di Sandy Bridge.

Verrà inoltre aggiunta nella GPU una cache L3 specifica per la grafica che permetterà l'accesso ai dati necessari per le elaborazioni grafiche diminuendo l'utilizzo della cache L3 della CPU ed evitando quindi di utilizzare il ring bus.[52]

[51] http://www.tomshw.it/cont/news/cpu-intel-ivy-bridge-sandy-bridge-non-terra-il-passo/29541/1.html
[52] http://www.anandtech.com/show/4830/intels-ivy-bridge-architecture-exposed/5

Per quanto riguarda il chipset, dovrebbe supportare nativamente, attraverso quattro porte**, l'USB 3.0**.

Al momento non sono previsti grandi cambiamenti in termini di funzionalità base ma solo una "preparazione" per il transito alla nuova tecnologia di produzione a 22nm per la nuova microarchitettura (**Haswell**) che vedrà la luce nel 2013.

§ 3.2 Processori di undicesima generazione (2013 - Haswell 22nm)

E' previsto l'impiego del processo produttivo a 22 nm introdotto con Ivy Bridge.[53]

Probabilmente i processori che saranno basati su tale architettura saranno tutti dotati di almeno 8 core a 14 stadi di pipeline, mentre per quanto riguarda la cache sarà basata su un progetto completamente innovativo; probabilmente sarà presente una L1 da 128 KB (64 KB per i dati e 64 KB per le istruzioni) con associatività a 4 vie, una L2 da 1 MB per ciascun core, sempre con associatività a 4 vie, e una L3 da 16 MB condivisa tra tutti i core con associatività a 8 vie.

Altre caratteristiche previste dalla nuova architettura dovrebbero essere un rivoluzionario approccio al contenimento dei consumi della CPU, fino alla metà rispetto a quelli attuali, grazie all'uso dei transistor 3D.

Dovrebbero essere presenti 2 ulteriori unità di calcolo inedite; la prima consiste nella possibilità di integrare nei futuri processori anche un coprocessore vettoriale che si occuperà dell'elaborazione di questo particolare tipo di calcoli, mentre dovrebbe fare il suo debutto anche un nuovo set di istruzioni Avx2 (Advanced Vector Extensions 2) con supporto per interi a 256bit e supporto per le istruzioni FMA3 (Fused Multiply-Add), che consentiranno di effettuare simultaneamente un'operazione di moltiplicazione e una di addizione attraverso una sola istruzione. Sono queste operazione usate principalmente per grafica professionale ad alta qualità e riconoscimento visivo.[54]

[53] http://it.wikipedia.org/wiki/Haswell_%28hardware%29
[54] Mark Buxton (Intel),13 Giugno 2011, http://software.intel.com/en-us/blogs/2011/06/13/haswell-new-instruction-descriptions-now-available/

§ 4 – Conclusioni

Nel corso degli ultimi 40 anni l'evoluzione tecnologica del microprocessore inventato da Intel ha permesso di passare dal 4004 del 1971 che conteneva 2300 transistor ed era fabbricato con tecnologia a 10000nm all'attuale microprocessore Core 2nd generation di classe Sandy Bridge costituito da circa 1 miliardo di transistor con tecnologia a 32nm .

Rispetto al 4004, una CPU a 32nm è 4.000 volte più veloce e ogni transistor consuma 4000 volte meno energia senza contare il fatto che il prezzo per transistor è diminuito di un fattore 100.000.[55]

Nel corso di questi 40 anni sono emersi vari fattori che hanno contribuito a questo eccezionale risultato e da una iniziale ricerca di integrazione circuitale sempre più spinta e di un aumento della velocità di clock si è passati alla ridefinizione della geometria del transistor da planare a 3D e alla gestione dei consumi e del risparmio energetico.

Allo stesso modo l'architettura, pur mantenendo la compatibilità con la generazione x86, ha avuto una evoluzione ugualmente impressionante con la progressiva integrazione sul die della cpu di quasi tutti i chip necessari per gestire l'elaboratore a partire dalle varie memorie cache, del sottosistema di comunicazione con le periferiche fino ad arrivare alla inclusione della cpu grafica che rappresenta l'ultimo passo per gestire al meglio l'odierna mole di contenuti multimediali presenti in rete e non solo.

L'unica cosa certa è che la legge di Moore (che non è una legge vera e propria ma solo una osservazione empirica di un trend) che stabilisce un raddoppio dei componenti costituenti un microprocessore ogni 18 mesi probabilmente dovrà essere rivista in maniera migliorativa in un prossimo futuro perché con gli enormi investimenti le società produttrici di circuiti integrati riescono a sviluppare tecnologie fino a qualche anno fa impensabili che spostano sempre più nel futuro un non meglio identificato limite fisico. A tal proposito Intel stima che fra 10 anni un microprocessore potrà contenere anche 100 miliardi di transistor.

[55] http://www.techspot.com/guides/357-fun-facts-intel-sandy-bridge/

Come mostra la fig. che segue possiamo notare che in un futuro molto prossimo sono previste tecnologie come quella dei nanotubi in carbonio e delle interconnessioni ottiche che possono proiettare la tecnologia del microprocessore su piani di sviluppo che avranno evoluzioni al momento non ancora chiare ma sicuramente esaltanti da un punto di vista tecnico e computazionale.

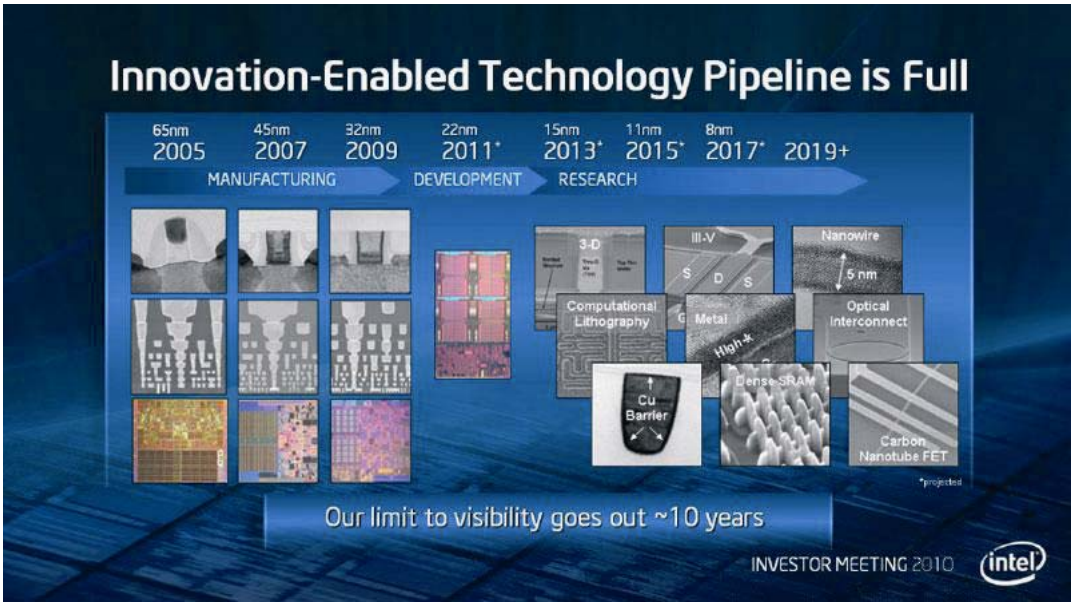

Figura 40 Roadmap Intel fino al 2019 [56]

[56] http://www.tomshw.it/cont/news/cpu-intel-ecco-la-roadmap-completa-fino-al-2019/25284/1.html