Università degli Studi di Padova

FACOLTÀ DI INGEGNERIA
Corso di Laurea Triennale in Ingegneria dell'Informazione

Tesi di laurea triennale

AMD Fusion

Prendono forma le prime APU

Candidato:
Giuseppe De Rito
Matricola 609596

Relatore:
Sergio Congiu

Anno Accademico 2012–2013

Indice

1 Da dove ha avuto origine il progetto

La nascita dei processori Fusion fonda le sue radici nella storia di due colossi dell' informatica: ATI e AMD.

- ATI Technologies è un' importante azienda canadese produttrice di GPU con sede a Markham, in Ontario. Fondata nel 1985, era riuscita, sotto l'abile guida di Dave Orton (presidente e direttore operativo), a raggiungere le prime posizioni nel mercato delle schede grafiche. L'azienda ottenne rapidamente una posizione di leadership nel settore, arrivando, nel 2002, a dominare la scena con le nuove schede grafiche compatibili con l' architettura Northbridge.

- Advanced Micro Devices (AMD) è invece una multinazionale americana produttrice di semiconduttori con sede a Sunnyvale, in California.
  Fondata da Jerry Sanders nel 1969 è la principale concorrente di Intel nel mercato dei microprocessori.

Fu proprio nel 2002, che Senders, dopo aver guidato l'azienda per ben tre decenni, ne lasciò la guida al suo braccio destro Hector Ruiz che divenne quindi il nuovo amministratore delegato di AMD.

Negli stessi anni, grazie ai processori Athlon, AMD riuscì per la prima volta a scavalcare Intel, fino ad allora leader indiscussa, e a superare per prima la soglia di 1 GHz di frequenza.
I processori Athlon ottennero un successo strepitoso, sia nella versione K7 del 1999 sia con la versione K8 del 2003 (che ebbe molta diffusione nel settore dei server).

Forte di questi successi, la casa di Sunnyvale conobbe un periodo di rapida crescita fino a Luglio del 2006 quando Intel svelò al pubblico l'architettura Core passando di nuovo in testa alle vendite.

Ormai già da qualche tempo Ruiz aveva capito che bisognava cambiare strada per poter prendere Intel in contropiede e nominò Dirk Meyer (allora responsabile del progetto Athlon) nuovo "capo operativo del settore microprocessori".
A quel punto i due cominciarono a studiare una nuova strategia da adottare e giunsero a conclusione che la vecchia architettura non era più sufficiente! Bisognava allargare i propri orizzonti di mercato, soprattutto nei segmenti mobile e consumer, acquisendo maggiore competitività nei confronti della rivale Intel.

Il culmine di questa strategia venne rivelato il 24 Luglio del 2006 da Hector Ruiz durante una conferenza tenutasi a New York.

AMD avrebbe acquistato ATI Technologies per un valore di ben 5,4 miliardi di dollari.

Figura 1: Dave Orton and Hector Ruiz

Il motivo di tale investimento risiedeva nella possibilità per AMD di metter mano su una grande quantità di proprietà intellettuali nel campo delle schede grafiche e di conseguenza poter avviare un nuovo grande progetto: fondere CPU e GPU su un unico die di silicio. Tale progetto venne chiamato appunto "Fusion" e venne pubblicizzato con il nome in codice "Llano". L'approccio a questa nuova strategia progettuale era dovuto al fatto che negli ultimi anni il mercato dei desktop-computer aveva iniziato una fase discendente mentre vedeva un aumento delle vendite di notebook e dispositivi mobili di vario genere.

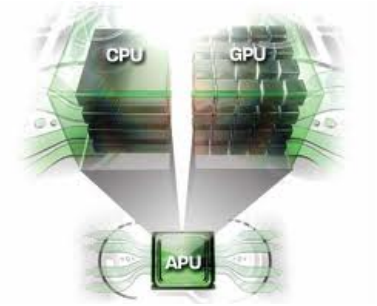

Figura 2: APU = CPU + GPU

La fusione di CPU e GPU in un unico chip, definita APU (Accelerated Processing Unit), avrebbe portato ad un risparmio di superficie oltre a notevoli consumi di energia con conseguente maggior durata delle batterie.

Inoltre tutto ciò significava un ampliamento del business verso la così detta elettronica di consumo (telefoni cellulari, decoder per televisioni digitali, console per videogiochi, navigatori satellitari ecc...).

2 Il lavoro ha inizio: i primi problemi

Come succede per ogni grande progetto, anche "Fusion" portò con sè molti problemi. In primis quelli economici.
E' inutile negarlo, l'acquisto di ATI fu un investimento molto rischioso per un'azienda come AMD. Basta pensare che al momento dell'acquisto il valore complessivo di AMD si attestava intorno ai 9 miliardi di dollari. Spenderne 5,4 per acquistare ATI costituiva senza dubbio un azzardo che richiedeva molto coraggio.

Figura 3: AMD Price history

Ad aumentare le polemiche riguardanti tale decisione contribuì il fatto che, almeno per i primi tempi, AMD, per salvaguardare le future vendite, non rivelò al pubblico quali fossero le sue reali intenzioni. Inoltre un progetto così ambizioso e delicato non era facile da spiegare e in particolare era difficile prevedere come sarebbe andata a finire.

La situazione economica dell'azienda si aggravò perché, oltre alla crisi economica di quegli anni, ci fu il ritardo nell'uscita delle schede grafiche

Radeon X1000 nel 2005 (la linea di produzione Radeon era stata avviata da ATI nel 2000) e le scarse novità introdotte dalle Radeon Hd 2000 che arrivarono tra l' Aprile e il Maggio del 2007.
Le prestazioni offerte dalla tecnologia Radeon non erano paragonabili a quelle offerte da NVidia (storica rivale di ATI nella produzione di GPU) con il modello GeForge nelle due versioni del 2005 e del 2006. Tutto ciò fece vacillare la fiducia dei mercati.

Basti pensare che all'inizio del 2006 il valore delle azioni AMD era intorno ai 40$ mentre appena un anno dopo era sceso a 2$.

Un altro aspetto che rallentò i lavori fu di tipo puramente filosofico.
All'interno del team di sviluppo di "Fusion" si erano creati due differenti correnti di pensiero riguardo a come avrebbero dovuto funzionare questi nuovi chip.

- Alcuni sostenevano che la GPU dovesse essere usata principalmente per elaborare i contenuti grafici, esattamente come avveniva con i chip separati;

- Altri volevano sfruttare una parte della potenza di calcolo della GPU per alleggerire il carico di lavoro della CPU. Però, togliere risorse alle funzioni proprie della GPU avrebbe avuto un impatto negativo sulle prestazioni grafiche.

Se ben gestita, la combinazione di CPU e GPU è di straordinaria potenza perché le CPU hanno un numero di core contenuto e sono ottimizzate per l'elaborazione seriale, mentre le GPU hanno migliaia di core più piccoli ed efficienti progettati per l'elaborazione in parallelo. Le porzioni seriali del codice vengono eseguite con la CPU mentre le porzioni parallele vengono eseguite dalla GPU.
Il motivo di tale differenza risiede nei diversi tipi di calcolo che GPU e CPU devono eseguire.

La GPU si occupa principalmente dell'accelerazione 3D - Quando viene elaborata un'immagine 3D, la CPU si occupa solo del calcolo delle coordinate geometriche dei vertici dei poligoni che compongono gli oggetti della scena e lascia alla GPU il compito di riempire le "facce" formate da questi vertici e il calcolo delle ombre e degli effetti grafici da applicare ai poligoni, sgravandosi da pesanti operazioni di calcolo. Tali operazioni sono, per loro natura, di tipo altamente parallelo e beneficiano ampiamente dell'architettura tipica delle GPU.

Oltre alle operazioni sulla grafica 3D la GPU entra in funzione anche nella maggior parte delle attività visive come la riproduzione di filmati HD e l'editing di video e foto.

Alcune operazioni generalmente eseguite dalla CPU si adattano anche ad essere eseguite dalla GPU e quindi è possibile spostare la loro esecuzione da un' unità di calcolo all'altra.

Bisognò quindi trovare un giusto equilibrio tra le due proposte e questo dibattito richiese quel tempo che i mercati non furono disposti a concedere!

L'ultimo e anche il più difficile dei problemi che si dovette affrontare nell'effettiva realizzazione dei processori Fusion fu di tipo tecnologico.

Far coesistere CPU e GPU sullo stesso die non era semplice perché le tecnologie fino a quel momento usate per la costruzione dei due processori erano diverse.

La CPU ha bisogno di transistor molto veloci e a bassissima resistenza; la GPU contiene molti più transistor che lavorano a una velocità ridotta.

Gli ingegneri che lavoravano per risolvere questo problema si resero conto che una soluzione ibrida non poteva essere realizzata con il processo produttivo a 45nm che allora veniva utilizzato.

A differenza delle CPU, le GPU utilizzano molti più transistor dedicati all'architettura rispetto a quanti ne sono impiegati in una CPU classica (in queste, molto spazio è occupato da transistor dedicati alle cache di secondo e terzo livello). Ciò comporta un consumo energetico elevato che richiede soluzioni specifiche di alimentazione e di raffreddamento.
Con la riduzione del processo produttivo si ottengono prestazioni superiori, un consumo sotto carico inferiore e uno a riposo capace di scendere alla soglia di pochi Watt.
Bisognava quindi attendere l'arrivo dei 32nm. Infatti questo processo è cresciuto di pari passo al progetto Fusion e fu reso disponibile sul finire del 2010.

La tecnologia, usata fin dal 2001 da AMD, per la produzione delle CPU era la IBM SOI (Silicon On Insulator) che consente di ridurre le capacità parassite tra le regioni attive e il bulk dei transistor inserendo tra essi uno strato di ossido isolante. Per poter ottimizzare il processo SOI a 32nm in modo che potesse andar bene per entrambe le architetture ci volle un anno di lavoro.

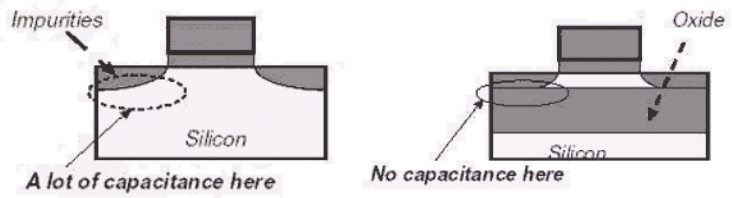

Figura 4: Difference between normal transistors and SOI transistors

3 Com'è costruita l'APU nel dettaglio

Il modello LLANO delle APU AMD presenta una struttura complessa (processore, unità grafica e circuiti integrati di vario genere) che permette agli utenti di ottenere ciò che loro veramente desiderano:

- migliore performance per le varie tipologie di applicazioni;
- aumento dell'autonomia della batteria;
- bassi costi.

In Llano questa integrazione non significa solo un'incorporazione fisica dei componenti su un unico die monolitico ma anche una gestione dinamica e sofisticata delle varie operazioni in modo da evitare conflitti tra i dispositivi. Sono stati combinati quattro processori x86 cores, un Unified Video Decoder, un comparto grafico compatibile con Direct X 11 (collezione di API per lo sviluppo di grafica avanzata) e tanti altri componenti.
Il chip ha un'area di 227 mmq con tecnologia a 32 nm SOI.

I quattro cores sono un' evoluzione dell'architettura AMD Stars a 45 nm, la stessa impiegata negli Athlon e nei Phenom II. Ognuno di questi cores presenta una memoria cache multi livello di 1Mbyte.
Le cache più grandi sono lente ma hanno un "hit-rate" molto alto in quanto possono memorizzare molti dati. Quelle piccole sono più veloci ma contengono meno dati e quindi è più facile incorrere in un "cache-miss".

Per ottenere le prestazioni massime, AMD ha fatto ricorso a una cache a due livelli (L1 ed L2):
quando il processore ha bisogno di accedere alla memoria, per prima cosa controlla se i dati sono presenti nella cache di primo livello (che essendo più piccola presenta tempi di risposta molto brevi) e, se non li trova, li va a cercare nel secondo livello (che è costituito da una memoria più grande e

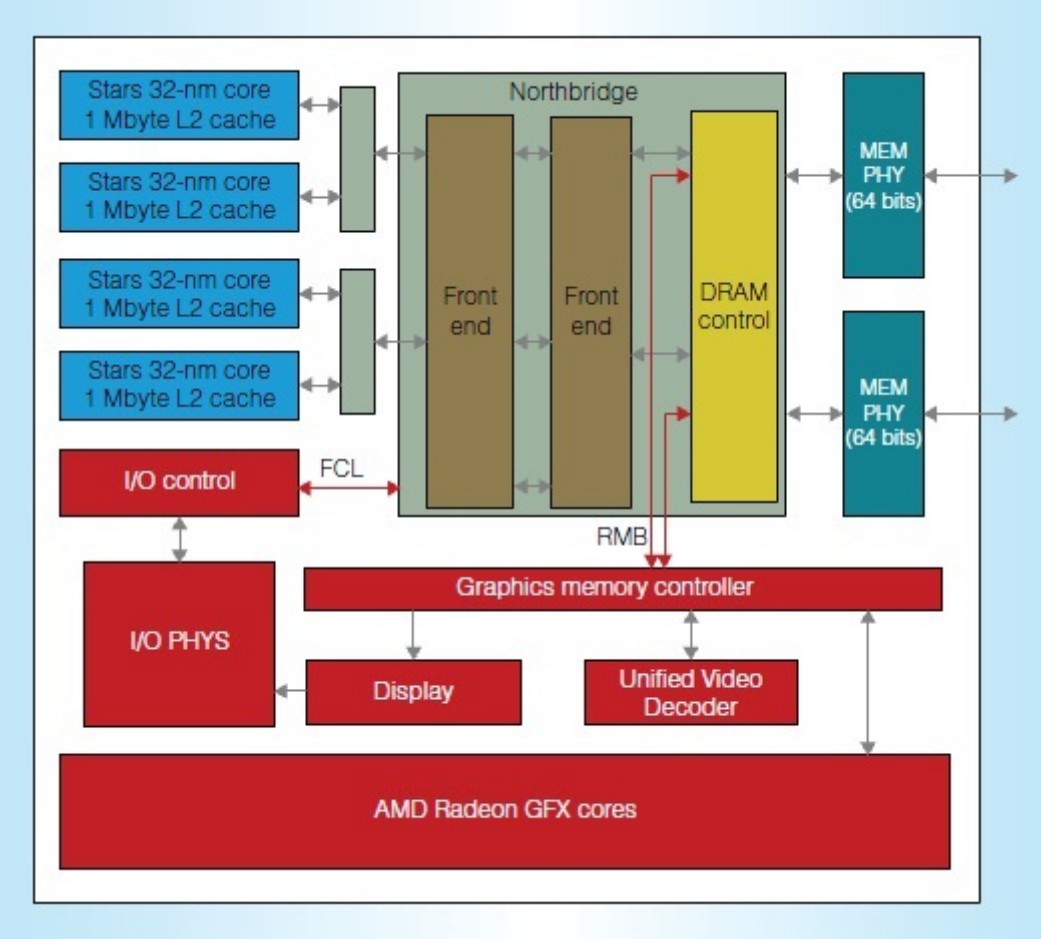

Figura 5: Llano accelerated processor unit (APU) block diagram. The block diagram shows the main APU components as well as links and busses for their interaction and access to main memory. (FCL: Fusion Control Link; MEM: memory; PHY/PHYS: physical layers; RMB: Radeon Memory Bus.)

più lenta ma che contiene più dati). Solo se il dato non è presente in nessuno dei due livelli bisogna andarlo a cercare nella memoria centrale. In questo modo vengono minimizzati i tempi di accesso.

A differenza di Intel che usa una memoria cache "inclusiva", AMD usa una memoria "esclusiva".
La differenza consiste nel fatto che i dati presenti nella memoria L1 non sono replicati anche nella memoria L2 (cosa che invece succede nella cache inclusiva).
Questo permette di memorizzare in complessivo un numero di dati maggiore e quindi avere un "hit-rate" più alto.

Un altro fattore che contribuisce a migliorare la performance significativamente è l'introduzione dell'instruction-level-parallelysm (ILP) e del memory-level-parallelysm (MLP).
L'ILP è un sistema che consente di eseguire più istruzioni contemporaneamente all'interno dei vari cores. L'MLP invece consente di effettuare più accessi in memoria allo stesso tempo.

Per consentire una miglior efficienza nell'esecuzione parallela delle istruzioni è stato inserito un Reordering Buffer (ROB) che permette l'esecuzione fuori ordine delle istruzioni ovvero senza rispettare l'ordine imposto dal programmatore.
Il processore analizza il codice che dovrà eseguire e individua le istruzioni che non sono vincolate dalle altre e le esegue in parallelo in modo da far lavorare tutte le unità funzionali, anche quelle che altrimenti rimarrebbero inutilizzate.
Il Reordering Buffer tiene traccia dell'ordine reale delle istruzioni, a mano a mano che queste vengono eseguite dalle unità funzionali, preleva i dati elaborati e li memorizza nei registri del processore seguendo l'ordine logico del programma.
Una volta completata questa procedura l'istruzione viene cancellata dal buffer.

Un altro compito svolto dal ROB è quello di garantire che l'esecuzione logica dei programmi venga preservata anche nel caso di errori di predizione dei "salti" (dovuti a cicli o condizioni) permettendo di eliminare le istruzioni eseguite erroneamente.

Allo scopo di aumentare il numero di istruzioni eseguite per ciclo c'è un instruction-pointer-based prefetcher. Tale dispositivo serve per accelerare l'esecuzione dei programmi eliminando i tempi di attesa necessari per il caricamento dei dati.

E' risaputo che l'operazione più lenta che pregiudica la velocità del processore è l'accesso in memoria.
Il prefetcher si occupa di caricare di volta in volta l'istruzione successiva durante l'esecuzione di quella corrente.

LLANO fa uso di due sistemi per ottimizzare il consumo di potenza e migliorare l'efficienza della batteria:

- power gating;
- dynamic voltage/clock frequency scaling.

Il primo è una tecnologia sviluppata per la prima volta da Intel che gestisce individualmente l'attività di ogni singolo core in base alle necessità del momento, spegnendoli o accendendoli quando necessario.
Il secondo permette di regolare individualmente la frequenza operativa dei core in base al carico di lavoro.

Per evitare gli sprechi energetici, ogni core implementa un sistema CAC di "digital monitoring of internal activity".
Questo sistema di monitoraggio permette di misurare costantemente il livello di attività dei singoli core e invia i dati al PMC (power management controller) che a sua volta li elabora e scala la frequenza operativa e la tensione in base ai carichi di lavoro.

Se non è necessario mantenere accesi tutti i core interviene il sistema di "power-gating" che li isola, insiema alle relative cache, dal resto del die.
Le linee di collegamento tra i core e gli altri elementi sono formate da transistor molto larghi che garantiscono una bassa resistenza.
Oltre ai core, queste due tecnologie sono state applicate anche alle altre unità di calcolo presenti nel chip (per esempio GPU e UDV).

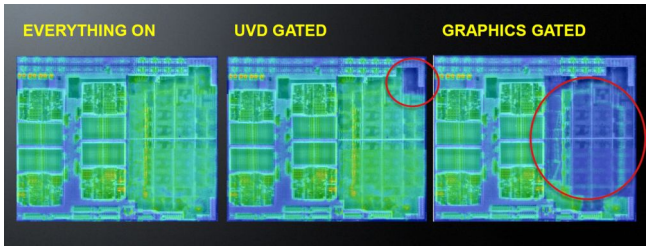

Figura 6: Llano al microscopio termico, completamente funzionante o, qualora l'applicazione non la richieda, con la sezione UVD spenta e con la grafica disattivata.

I quattro core comunicano con il resto dell'hardware veloce mediante il Northbrige (anche detto 'memory controller hub'). Questo costituisce una delle componenti fondamentali dell'intero sistema perché dalla sua velocità dipendono i tempi di trasmissione dei dati tra CPU e GPU e i tempi di accesso alla memoria. Inoltre presenta alcuni sistemi di controllo della tensione, mantenendola nei limiti fisici dei vari dispositivi. La sua frequenza operativa ha ricadute sulla performance dell'intero sistema e impone la durata del ciclo di clock.
Il Northbrige è collegato al Southbrige che a sua volta gestisce la comunicazione con l'hardware lento (periferiche esterne).

All'interno del Northbrige sono contenuti due unità Front-End, responsabili dell'acquisizione dei dati in ingresso e della loro elaborazione per renderli utilizzabili agli altri dispositivi della rete. Inoltre è presente il DRAM controller per la gestione della memoria condivisa (mediante allocazione di banda) permettendone l'accesso ai dispositivi secondo un ordine di priorità. Esso contiene tutte le funzioni logiche necessarie per la lettura e la scrittura della DRAM.
Se la DRAM non subisce cilci di refresh periodici, tutti i dati memorizzati vengono persi e questo è un altro dei compiti svolti dal controller.

Per quanto riguarda il comparto grafico, queste APU integrano un core di GPU DirectX 11 con tecnologia Radeon che supporta il sistema di power-gating esattamente come i core della CPU.
Se l'unità di controllo della potenza si accorge che la GPU non viene utilizzata per più di un certo intervallo di tempo, le varie informazioni elaborate vengono salvate e l'unità grafica viene spenta.
Il tempo necessario per spegnere o riattivare il core è dell'ordine di poche unità di microsecondo e quindi la transizione da uno stato all'altro non causa perdita di performance.

Per velocizzare l'esecuzione delle operazioni viene usato un sistema VLIW-5 (Very Long Instruction Word).
Sempre per lo stesso motivo la GPU ha accesso diretto alla memoria mediante il Radeon Memory Bus (RBM) a 256 bit.

La potenza complessiva dell'unità di calcolo si attesta intorno ai 480 Gflops (flops = floating point operations per second).

Affiancato alla GPU c'è un Unified Video Decoder di terza generazione (UVD3) che viene impiegato nella decodifica hardware dei codec video.
Tra le codifiche supportate ci sono: Multi-View-Codec (codec di compressione di sequenze video catturate da più videocamere simultaneamente usando un unico flusso video con un risparmio del 50% di

memoria per ogni traccia supplementare), MPEG-4, MPEG-2, Blu-Ray e HD-DVD.

Llano utilizza anche un "unified memory architecture" (UMA) nella quale il processore e la grafica condividono i dati. Una parte di questa memoria viene dedicata ad un frame-buffer dove vengono memorizzate informazioni relative ai fotogrammi da visualizzare.
Il flusso di dati tra la memoria e l'unità grafica viene gestito da un dispositivo chiamato "Graphics Memory Controller" (GMC).

L' I/O Controller invece ha un accesso diretto al Northbridge e quindi alla memoria mediante il Fusion Control Link (FCL) a 128 bit. Questo dispositivo è incaricato di ricevere le richieste di input/output dal processore e di inviare alle periferiche esterne specifici segnali di controllo. Tutto ciò serve per liberare il processore dal compito di dover controllare ogni singolo dispositivo.

Infine è presente un Display Controller che conserva i frame pronti per essere inviati all'output.

Per massimizzare le prestazioni dell'APU nei limiti del TDP (Thermal Design Power = valore indicativo del calore dissipato dal processore e che deve essere smaltito dal sistema di raffreddamento riferito a condizioni di utilizzo "normali") è stata adottata la tecnologia AMD Turbo CORE:
se l'attività dei singoli core è tale da mantenere il TDP sotto la soglia stabilita, il sistema incrementa automaticamente la frequenza di alcuni core mandandoli in "overclock".
Quando il carico di lavoro di un core è eccessivo rispetto ad un altro, è possibile aumentare la frequenza del core in questione e diminuire quella dell'altro mantenendo comunque un valore complessivo del calore intorno alla media.

L'unità addetta alla misurazione del TDP è l' APM, ovvero l' Advanced Power Management.
A differenza dei processori Intel, il calcolo del TDP non avviene in base a misurazioni reali della temperatura nelle singole parti del circuito ma viene stimata in base ad alcuni valori attribuiti a priori alle singole attività svolte (infatti non tutte le operazioni svolte dal processore producono la stessa dissipazione di calore).
Purtroppo la GPU non può essere overclockata perché la sua frequenza operativa standard è già quella massima imposta dalla tecnologia e di conseguenza è solo possibile diminuirla.
Quindi nel modello LLANO possono essere overclockati solo i core della CPU (nelle recenti APU Trinity è possibile overcloccare anche la GPU).

Se CPU e GPU sono entrambe sovraccariche la precedenza viene sempre data alla GPU.

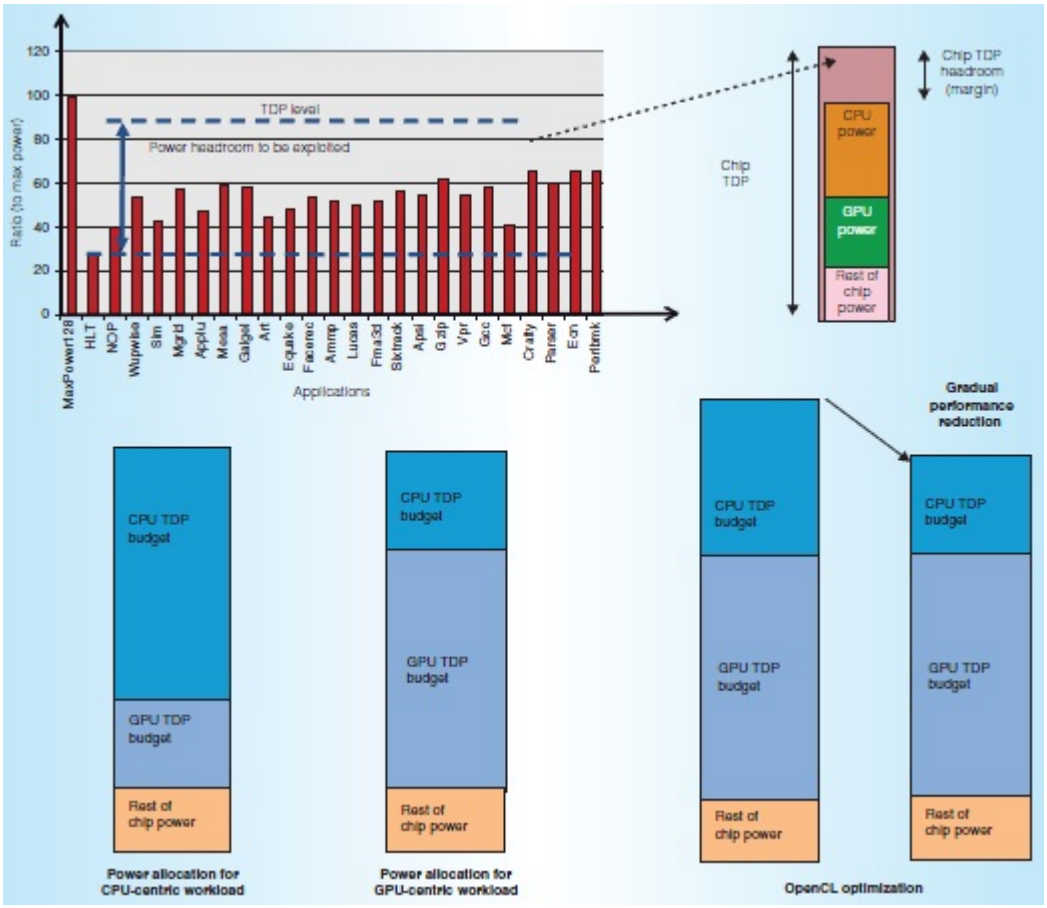

Figura 7: AMD Turbo CORE concept in Llano. Different application classes can use available chip TDP headroom to deliver higher performance.

In realtà, in condizioni di lavoro "normali" viene sempre lasciato un margine tra la soglia massima di TDP raggiungibile e quella realmente utilizzata. Tale margine viene chiamato "Power Headroom" e può essere utilizzato a favore delle varie unità non appena il carico di lavoro comincia a farsi pesante.
In llano sono implementati due diversi tipi di "Energy Margin Accumulator":

- chip level margin accumulator: indica il margine relativamente all'intera APU;

- CPU energy margin accumulator: indica il margine relativamente ad ogni singolo core del processore.

Se in un determinato istante il margine di energia complessivamente utilizzabile è molto grande, il chip level margin accumulator impedisce che il TDP del singolo core possa essere spinto erroneamente oltre i limiti tecnologici quando tutta la potenza viene messa a sua disposizione.

Bisogna inoltre considerare che la capacità di raffreddamento dei core varia in base alla situazione dei core adiacenti. Infatti se i core adiacenti non stanno lavorando, parte del calore può essere disperso lateralmente. Se però i core vicini sono anch'essi a pieno regime e quindi producono calore, non è più possibile sfruttare la superficie laterale per far abbassare la temperatura e quindi bisogna limitare la produzione di calore per evitare guasti.

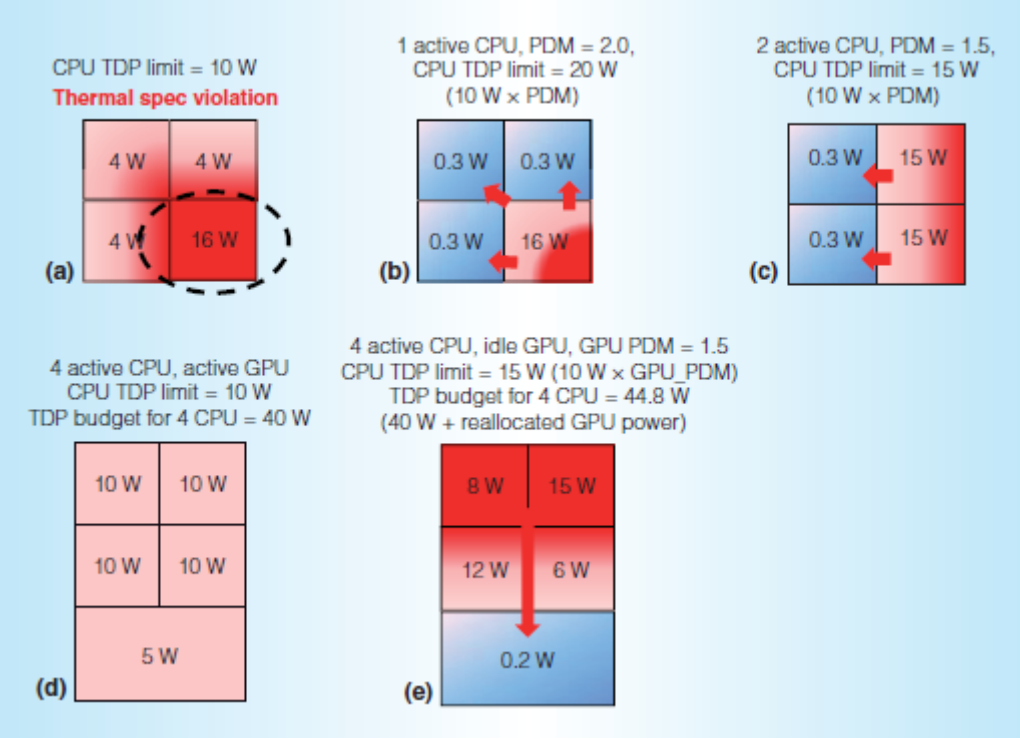

Figura 8: AMD Turbo CORE management of idle compute units.

Proprio per questo motivo è stato introdotto il concetto di TDP variabile. Per far fronte a questo problema viene applicato al TDP dei vari core un fattore di moltiplicazione: il Power Density Multiplier (PDM). Variando il

PDM è possibile regolare dinamicamente il valore massimo del TDP di volta in volta.
Se ad esempio un core ha bisogno di aumentare la propria frequenza operativa è sufficiente aumentare il suo PDM per mettere a sua disposizione un margine di TDP maggiore.
Se invece i core vicini stanno producendo molto calore e abbiamo bisogno di abbassare la temperatura dell'intera zona è sufficiente abbassare il TDP dei core e quindi abbassare il loro PDM.

Lo stesso concetto può essere esteso anche alla GPU.
Infatti anche il TDP della GPU può essere regolato mediante un apposito PDM.
Quando la GPU non viene usata, la sua frequenza viene ridotta e in questo modo i core possono sfruttarla per disperdere calore.

Questi valori del TDP sono studiati in modo tale da massimizzare le prestazioni sia della CPU che della GPU e ottenere quindi il miglior rapporto tra potenza dissipata e performance.
L'ottimizzazione è stata accuratamente studiata anche nei casi di sovraccarico. Per soddisfare questo requisito viene sempre allocato più TDP alla GPU rispetto alla CPU perché è dimostrato che a parità di potenza dissipata la GPU garantisce prestazioni superiori (in termini di Gflops/watt).

4 Finalmente sul mercato

AMD ha mostrato le sue prime APU Fusion al CES (Consumer Electronics Show) del 2011 e subito dopo ne ha iniziato la distribuzione con le soluzioni per il settore consumer: A4, A6, A8 ed E2.

Un altro annuncio fatto al CES era che la Fusion System Architecture sarebbe diventata Heterogeneous System Architecture (HSA). Questa modifica nella nomenclatura stava ad indicare la nascita di un nuovo standard industriale aperto.
L'obbiettivo di AMD non era quello di sviluppare un progetto fine a se stesso ma di cambiare radicalmente la rotta nella realizzazione dei computer moderni.

Almeno per il momento, la grafica integrata non è in grado di competere con quella dedicata nei computer di fascia alta ma in compenso offre un ottimo rapporto qualità prezzo e quindi si adatta molto bene ad essere adottata su computer mainstream.
Tutto ciò non significa che il suo impiego sia limitato ma anzi il contrario.

Processori AMD "Llano" desktop												notebook
Modello	A8- 3560P	A8- 3550P	A8- 3560	A8- 3550	A6- 3460P	A6- 3450P	A6- 3460	A6- 3450	A4- 3360	A4- 3350	E2- 3250	A8-3510MX
Core	Husky	Husky	Husky	Husky	Husky	Husky	Husky	Husky	Husky	Husky	Husky	Husky
Numero core	4	4	4	4	4	4	4	4	2	2	2	4
Stream processor	400	400	400	400	320	320	320	320	160	160	160	400
Frequenza Cpu	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	1,8 GHz
Frequenza Gpu	n.d.	594 MHz	n.d.	594 MHz	n.d.	443 MHz	n.d.	443	n.d.	594 MHz	443 MHz	n.d.
Gpu integrata	HD 6550	HD 6550	HD 6550	HD 6550	HD 6530	HD 6530	HD 6530	HD 6530	HD 6410	HD 6410	HD 6370	HD 6620
Cache L2	4 MB	4 MB	4 MB	4 MB	4 MB	4 MB	4 MB	4 MB	2 MB	2 MB	1 MB	4 MB
Memoria Ddr3	1.866 MHz	1.866 MHz	1.866 MHz	1.866 MHz	1.866 MHz	1.866 MHz	1.866 MHz	1.866 MHz	1.866 MHz	1.866 MHz	1.600 MHz	1.333 MHz
Tecnologia	32nm	32nm	32nm	32nm	32nm	32nm	32nm	32nm	32nm	32nm	32nm	32nm
TDP	100 W	100 W	65 W	65 W	100 W	100 W	65 W	65 W	65 W	65 W	65 W	45 W

Figura 9: LLANO processors.

Infatti secondo le stime fatte da IDC (International Data Corporation), già verso la fine del 2011, 3 computer su 4 montavano grafica integrata.
Inoltre non bisogna limitarsi al settore dei computer ma bisogna estendere la propria visione a tutti i dispositivi dell'elettronica di consumo. E' proprio in questo settore che le APU possono esprimersi al meglio.
Se ci riferiamo a questo mercato allargato, la metà di tutti i processori venduti, compresi quelli degli smartphone, sono APU.
E' proprio nei dispositivi mobili che si fa sentire di più l'esigenza di avere alte prestazioni con consumi ridotti ed è proprio ciò che HSA permette di ottenere, ovviamente a patto che i programmi siano scritti in modo da sfruttare al meglio tali possibilità.
Purtroppo, nella maggior parte dei casi, quest'ultima condizione non è tuttora verificata.

5 Maggior attenzione al software

Fino a prima dell'arrivo di questa nuova tecnologia, il mondo della programmazione orbitava intorno alla CPU, e passare parte del codice alla GPU era tutt'altro che semplice.

Già in precedenza sia Nvidia che ATI avevano lavorato alle così dette GPGPU (General-Purpose GPU) nelle quali avevano tentato di sfruttare il processore della scheda grafica per scopi diversi dalla tradizionale creazione di un immagine tridimensionale.
Alcuni esempi di GPGPU erano le schede Stream di ATI o le ancor meglio riuscite schede CUDA di Nvidia.

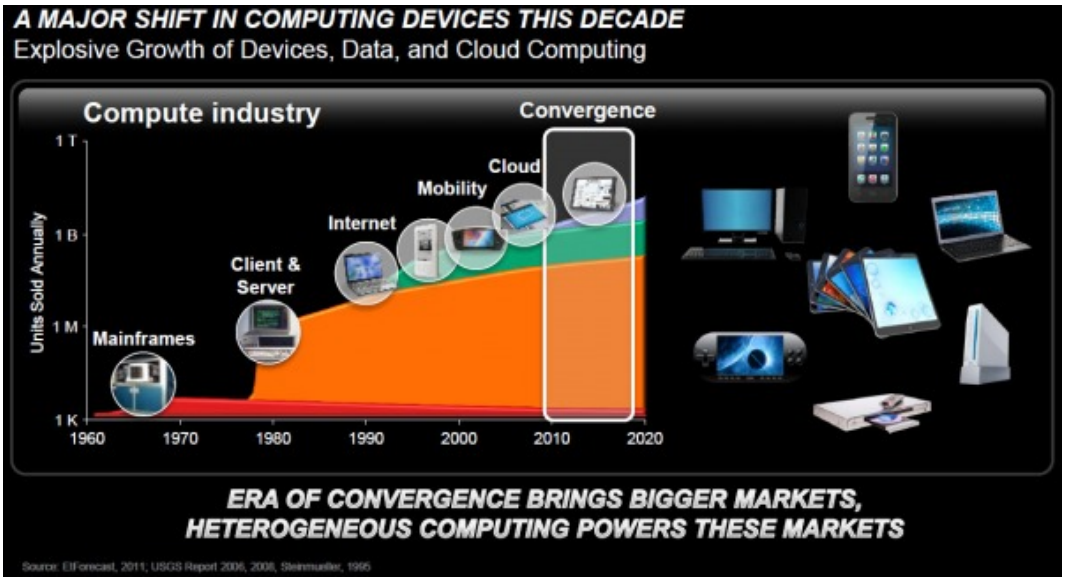

Figura 10: Market trends.

Purtroppo il calcolo GPGPU (da sempre sostenuto da AMD), nonostante le prestazioni superiori in alcuni settori come ad esempio nei giochi, aveva avuto grandi difficoltà nell'affermarsi più per le difficoltà di programmazione che per il fatto di essere un mercato di nicchia.
Il problema principale era quindi l'aumento del carico di lavoro sul programmatore, in termini di complessità del codice. Inoltre un'applicazione progettata per questo sistema era difficilmente trasportabile sulle altre piattaforme.
Bisognava riprogettare interamente le API tradizionali per interagire con la GPU. Queste erano state sviluppate per operazioni grafiche (gestione delle texture e dei poligoni) e non per i calcoli standard.
Questo passo era molto impegnativo ma allo stesso tempo necessario per permettere alle nuove APU di decollare (a differenza di come era successo per le GPGPU).

Venne quindi creato un nuovo standard, l' OpenCL (Open Computing Language) il cui scopo è quello di offrire al programmatore un'interfaccia semplice e diretta per sfruttare la GPU.
Questo standard, che viene gestito dal gruppo no-profit Khronos Group, ha trovato l'approvazione e il sostegno di molti grandi nomi del settore tra cui AMD, ARM, Intel, Nvidia e Apple.
OpenCL non si limita ad offrire le librerie necessarie per la programmazione

Figura 11: Logo OpenCL.

ma mette a disposizione degli sviluppatori anche l'ambiente software necessario.

L'utilizzo di piattaforme HSA presenta anche altri vantaggi per l'esecuzione delle applicazioni.
Ad esempio è stato risolto il problema delle comunicazioni fra CPU e GPU. Con questa nuova tecnologia le applicazioni possono mandare il lavoro direttamente al processore grafico e le due unità possono lavorare insieme allo stesso set di dati.
Si evita così di dover copiare e trasferire i dati tra le varie memorie.

La memoria condivisa rende tutto il sistema più facile da programmare ma questo non basta.
Per agevolare il lavoro dei programmatori bisognava rendere possibile l'utilizzo di linguaggi di alto livello. Ogni programmatore ha il suo linguaggio preferito e non è possibile chiedergli di cambiarlo.
HSA permette di sfruttare il calcolo eterogeneo per tutti i linguaggi di alto livello.
Infatti, oltre alla compatibilità con C e C++, AMD ha voluto assicurarsi una copertura totale e quindi ha fatto in modo che HSA funzionasse anche con C#, Java e persino con i linguaggi funzionali.

Fare tutto ciò era comunque un lavoro troppo grande per la sola AMD. Era quindi fondamentale trovare la collaborazione di altri partner.
E' proprio per questo motivo che HSA doveva necessariamente diventare uno standard aperto gestito da un' omonima associazione. Attualmente tra i partner di AMD vi sono ARM, MediaTek e Texsas Instruments.
La fondazione venne presentata al pubblico nel Giugno del 2012 con lo scopo di promuovere piattaforme e software compatibili con HSA. In poche parole bisognava creare SDK (Software Development Kit) per gli sviluppatori.
Tra gli strumenti sviluppati troviamo compilatori, debugger e librerie e sono

quasi tutti open-source in modo da minimizzare i costi di produzione del software per questa piattaforma.
Ovviamente queste librerie devono essere molto semplici da usare. Ad esempio, per i programmatori C++, è sufficiente usare la libreria C++ AMP (Accelerated Massive Parallelism C++) con la quale basta aggiungere due keyword al linguaggio ("restrict" e "array view") per rendere il programma pronto a gestire il carico tramite la GPU.

Infine, per incoraggiare lo sviluppo di applicazioni compatibili con HSA, AMD ha deciso di aiutare economicamente la nascita delle piccole aziende che decidono di sviluppare software secondo questo nuovo standard.

6 Il mercato secondo AMD

A questo punto viene spontaneo chiedersi quali siano le fasce del mercato alle quali punta AMD.
Questo emerge molto chiaramente da quanto è stato detto durante l'ultimo AMD Fusion Developer Summit (AFDS) tenutosi a Seattle. E' risultato evidente che il settore dei desktop non costituisce la priorità per l'azienda di Sunnyvale.
Infatti, in termini di vendite, sono i computer portatili a dettare legge mentre i desktop sono in stallo. Questo non significa che AMD abbia rinunciato in partenza al mercato dei fissi.

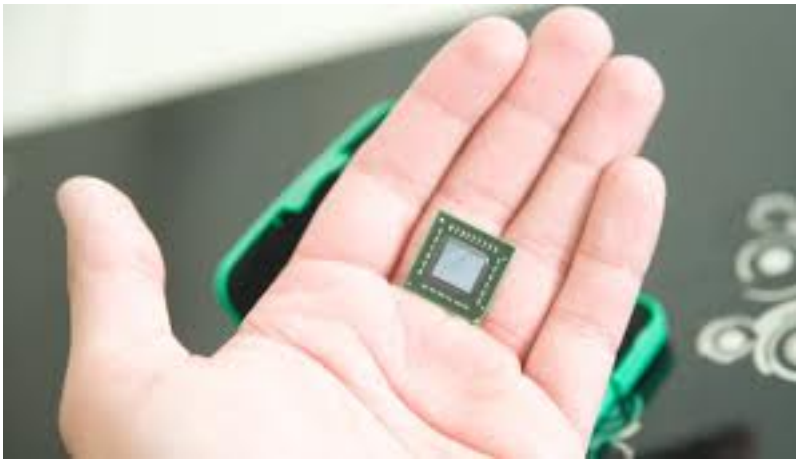

Figura 12: AMD Fusion.

Le nuove APU garantiscono, oltre ai bassi consumi e quindi un'ottimizzazione della batteria, anche silenziosità, sistema di

raffreddamento efficiente e alte prestazioni soprattutto nel settore video. Costituiscono quindi un'ottima scelta per i videogiocatori che vogliono avere una buona piattaforma di gioco a prezzo contenuto.

All'altro estremo del mercato ci sono i server e questi possono trarre parecchio vantaggio dall' HSA.
Nelle così dette "Server Farm" l'efficienza energetica è fondamentale, per non parlare dei sistemi di raffreddamento.

7 Sguardo al presente: le nuove APU Trinity

Le prime piattaforme AMD Trinity sono state annunciate il 15 Maggio di quest'anno e almeno per ora consistono esclusivamente in soluzioni per notebook (anche se quelle per pc arriveranno a breve).

La tecnologia produttiva è rimasta invariata a 32 nm (affidata alla GlobalFounderies) ma nonostante ciò le novità introdotte, con lo scopo di migliorare le prestazione e ridurre ulteriormente i consumi, non sono poche.

A differenza di Llano che montava una CPU derivata dagli Athlon, Trinity vede una CPU Piledriver di derivazione Bulldozzer. Si tratta quindi di due architetture completamente diverse.
Questo è dovuto al fatto che la tecnologia alla base di Llano aveva deluso per quanto riguarda la pura potenza di calcolo che non riusciva a competere con quella delle piattaforme concorrenti anche di fascia più bassa.
Infatti la capacità di calcolo della CPU è sempre stato il "tallone d'Achille" delle soluzioni AMD mentre per quanto riguarda le potenzialità grafiche si è trovata sempre più avanti rispetto a Intel.
E' stato quindi necessario abbandonare la vecchia architettura per puntare su qualcosa di innovativo.

Piledriver può essere considerata come la seconda generazione dell'architettura Bulldozzer introdotta da AMD nell'autunno del 2011.
All'interno di Trinity sono presenti due moduli Piledriver, ognuno dei quali può essere considerato come una CPU dual core.
Come in Llano manca la cache L3 mentre la L2 è unica per ognuno dei moduli ed è condivisa fra i core. Allo stesso modo vengono condivisi anche il fetcher e il decoder (componente elettronico che, in base alla combinazione di bit presenti ai suoi ingressi, attiva una o più linee di uscita).
Rispetto all'architettura Bulldozzer originaria sono state effettuate parecchie modifiche come ad esempio l'ottimizzazione delle unità di branch prediction (BPU), che si occupa di prevedere l'esito delle istruzioni di salto condizionato evitando gli eventuali rallentamenti, e del prefetcher.

Il buffer L1 Tlb (Translation Look-aside Buffer), è cresciuto di dimensioni. Questo componente è un buffer che la Memory Management Unit (MMU) usa per velocizzare la traduzione degli "indirizzi virtuali" in "indirizzi fisici". L' MMU è molto importante per i processori moderni in quanto permette di utilizzare la memoria centrale evitando interferenze tra i vari programmi in esecuzione o per far comunicare più processi tra loro nei sistemi multi-tasking.
Tale sistema sfrutta il fatto che la memoria indirizzabile dai bit del processore è di gran lunga superiore a quella realmente esistente (memoria fisica).
Per tradurre gli indirizzi virtuali in indirizzi fisici bisogna usare una "Page Table".
Quindi per ogni accesso in memoria virtuale bisogna effettuare due accessi in memoria fisica (prima per la page table e poi per leggere il dato cercato) e questo rallenta parecchio le prestazioni del processore. Per velocizzare le operazioni una parte di questa page table viene tenuta nella cache all'interno del Tlb. Pertanto avere un Tlb più grande permette di dover fare meno accessi in memoria.

Un'altra novità è che anche la GPU può variare automaticamente la frequenza di clock a seconda del carico di lavoro nei limiti del proprio TDP. Rispetto alla GPU di Llano che conteneva 1,17 miliardi di transistor, quella di Trinity ne contiene ben 1,3 miliardi. Di conseguenza sono cresciute anche le dimensioni del die che misura 246 millimetri rispetto ai 228 millimentri di Llano.

La GPU adotta un'architettura VLIW-4 (Very long instruction word) nota anche come Northern Island. A differenza della VLIW-5 presente in Llano, la VLIW-4 ha un'unità di calcolo in meno all'interno di ogni Thread-Processor (da 5 unità si è passati a 4 unità).
Questa scelta, secondo AMD, ha un impatto trascurabile sulle prestazioni poichè la quinta unità restava il più delle volte inutilizzata.

Una grande e importante novità di Trinity è il supporto alle porte USB 3.0, fino a 4, oltre alle 10 classiche USB 2.0. Questo è molto importante perché permette di lottare ad armi pari contro la concorrenza (Apple in particolare).

L'ultima differenza sostanziale che c'è tra Trinity e Llano sta nella piattaforma socket.
Il socket è un connettore elettrico che permette di fissare meccanicamente un circuito integrato sopra il circuito stampato (che a sua volta funge da collegamento tra i circuiti integrati).

Il nuovo socket FM2 differisce lievemente dall'FM1 di Llano e questo comporta che non è possibile istallare le nuove APU su una piattaforma progettata per il vecchio modello.

AMD serie A di seconda generazione - Trinity

Modello	TDP	Core	Clock	Clock Turbo	Cache L2	GPU	SP GPU	Clock GPU	Memoria DDR3	Prezzo
A10-5800K	100W	4	3,8 GHz	4,2 GHz	4MB	HD 7660D	384	800 MHz	1.866 MHz	122$
A10-5700	65W	4	3,4 GHz	4 GHz	4MB	HD 7660D	384	800 MHz	1.866 MHz	122$

AMD serie A di prima generazione - Llano

Modello	TDP	Core	Clock	Clock Turbo	Cache L2	GPU	SP GPU	Clock GPU	Memoria DDR3	Prezzo
A8-3870K	100W	4	3 GHz	-	4MB	HD 6550D	400	600 MHz	1.866 MHz	101$
A8-3850	100W	4	2,9 GHz	-	4MB	HD 6550D	400	600 MHz	1.866 MHz	91$
A6-3650	100W	4	2,6 GHz	-	4MB	HD 6530D	320	444 MHz	1.866 MHz	77$

Intel Core i3

Modello	TDP	Clock	Core	Thread	Cache L3	Unlocked	Turbo	GPU	Clock GPU	Turbo GPU	Prezzo
Core i3-3240	55W	3,4 GHz	2	4	3MB	no	-	HD2500	650 MHz	1.050 MHz	138$
Core i3-3220	55W	3,3 GHz	2	4	3MB	no	-	HD2500	650 MHz	1.050 MHz	117$

Figura 13: Caratteristiche a confronto: Trinity, Llano, Intel Core i3.

Dai test effettuati sulle varie piattaforme si è notato che il principale limite delle APU AMD continua ad essere il processo produttivo. Non tanto per la quantità di componenti integrati quanto per l'impossibilità di scendere ai livelli di consumo a pieno carico che invece si trovano nei processori Intel a 22 nm.
Il debutto di un nuovo processo produttivo per AMD è atteso entro il prossimo anno.

Il punto di forza è invece la grafica integrata che è di gran lunga superiore sia a quella di Llano che a quella delle schede integrate Intel.
C'è stato un netto aumento dei frame per secondo che si attesta intorno al 40% in più rispetto a Llano.
Nei confronti della GPU HD 4000 presentata da Intel come top di gamma il margine è ancora più consistente superando il 65%.

Per quanto riguarda la CPU si nota che la nuova architettura Piledriver non presenta grandissimi miglioramenti quando si lavora con un singolo

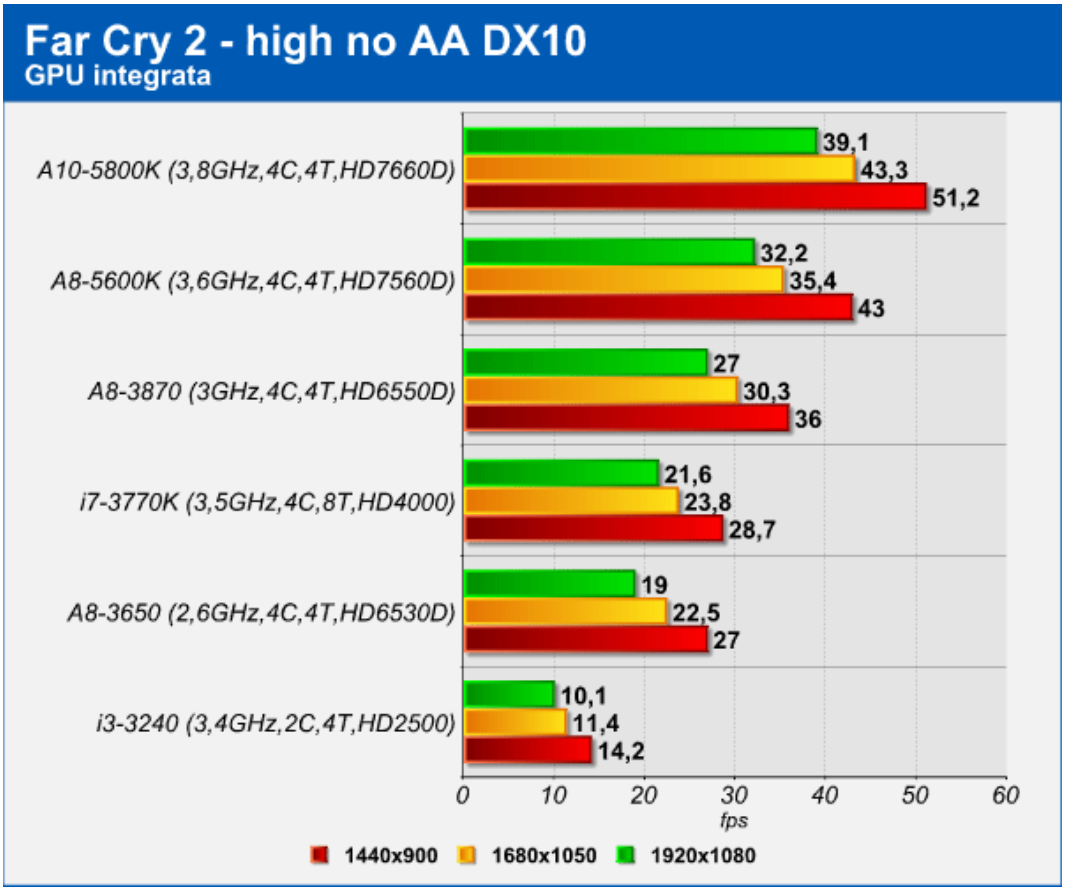

Figura 14: Test grafico effettuato sul videogioco Far Cry 2.

core. I vantaggi, rispetto a Llano, si notano solo quando il software è sviluppato al meglio per utilizzare tutti i core.

8 Aspettative per il futuro

Anche se il termine "chip integrato" viene spesso considerato sinonimo di basse prestazioni, non c'è nulla di più sbagliato. Infatti per poter valutare la reale efficienza di questa tecnologia bisognerebbe testarla nel suo ambiente naturale, cosa che oggi giorno non è ancora possibile fare.
In un ambiente di software sviluppato appositamente per sfruttare le caratteristiche delle APU, l'integrazione di CPU e GPU potrebbe offrire prestazioni di gran lunga superiore alle migliori piattaforme con scheda grafica separata.

Se si fa un paragone tra le attuali APU AMD e le CPU Intel, utilizzando i benchmark, si evince che queste ultime le superano nella pura potenza di calcolo. Le APU AMD vincono invece nella grafica (a parità di prezzo). Ma ciò è dovuto al fatto che le due aziende hanno puntato su soluzioni costruttive diverse, orientate a diverse fasce di mercato.
Intel punta al miglioramento dell'hardware spingendo le tecnologie produttive ai limiti della fisica.
Le ultime piattaforme Intel sono state realizzate con un processo fotolitografico a 22nm e sono già in fase di realizzazione prototipi a 14 nm.

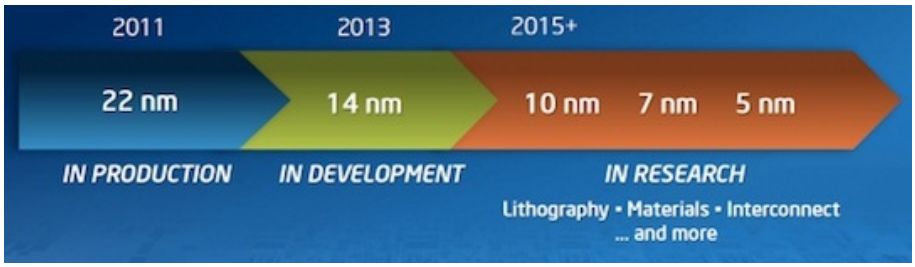

Figura 15: Intel: Innovation for next decade of computing.

Le più recenti APU Trinity di AMD (ultima evoluzione della tecnologia Fusion) invece sono realizzate ancora a 32 nm. AMD punta ad ottimizzare le tecnologie esistenti piuttosto che svilupparne di nuove.
Queste due strade hanno ognuna i suoi pregi e i suoi difetti ma sono senza dubbio due facce della stessa medaglia. Infatti l'obiettivo finale è praticamente lo stesso: ottenere prestazioni sempre superiori.

Il lavoro svolto da AMD potrebbe rivelarsi di fondamentale importanza in futuro.
Al giorno d'oggi la legge di Moore comincia a mostrare qualche crepa e costruire transistor più piccoli diventa sempre più difficile ed enormemente costoso. Per rendersi conto di quanto si è ormai vicini agli estremi della fisica basti pensare che, a 22 nm, lo spessore del gate di un transistor è formato da appena 10 atomi di silicio.
Potenziando meglio le tecnologie esistenti si potrà supplire all'inevitabile rallentamento che si avrà nello sviluppo dell'ulteriore miniaturizzazione dei processi litografici, consentendo ugualmente il miglioramento del trend tecnologico.

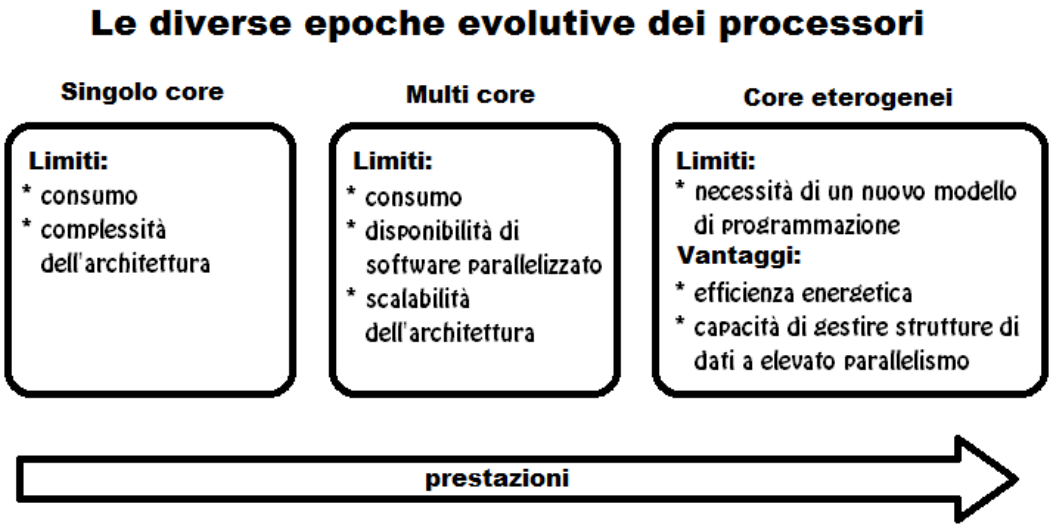

Figura 16: Evoluzione dei processori.

Da notizie di stampa sappiamo che i ricercatori sono già a lavoro su questo fronte. Si pensa di abbandonare definitivamente la forma storica dei transistor, modificando il design base che ci ha accompagnati sin dalla nascita della moderna elettronica.
Verranno aggiunti gate secondari a quello standard che dovrebbero migliorare i tempi di risposta dei segnali e verranno sviluppate costruzioni tridimensionali che garantiranno prestazioni notevolmente superiori.
I transistor del futuro non saranno più costruiti solo in silicio, ma si utilizzeranno altri materiali appartenenti al III e IV gruppo della tavola periodica, come il gallio, l'indio, il fosforo, l'arsenico o l'antimonio.

All'interno delle CPU verranno integrati nuovi componenti come le interconnessioni ottiche (il segnale si propaga mediante raggi di fotoni) o i nanotubi di carbonio (particolari strutture formate da atomi di carbonio che presentano un comportamento simile a quello del silicio presente nei microchip ma con prestazioni molto superiori) che spingeranno i processori verso potenzialità che sono ancora del tutto ignote.

Riferimenti bibliografici

[1] Alexander Branover, Denis Foley, Maurice Steinman *IEEE Computer Society*. AMD Fusion APU: Llano, 2012, Advanced Micro Devices (AMD).

[2] Michele Braga *PC Professionale*. La guida alle CPU, Marzo 2012, n.252.

[3] Michele Braga *PC Professionale*. Rivoluzione grafica, Febbraio 2012, n.251.

[4] Davide Piumetti *PC Professionale*. AMD Fusion: ecco i primi processori APU nome in codice Llano, Giugno 2011.

[5] La Redazione *PC Professionale*. AMD Fusion: i primi passi, Marzo 2011.

[6] Pasquale Bruno *PC Professionale*. AMD Trinity: APU di seconda generazione, Settembre 2012.

[7] William Van Winkle. *http://www.tomshw.it/cont/articolo/la-storia-di-amd-fusion-com-e-nato-cos-e-e-dove-ci-portera/39314/1.html*. La Storia di AMD Fusion: com'è nato, cos'è e dove ci porterà, 29 agosto 2012.

[8] Don Woligroski. *http://www.tomshw.it/cont/articolo/apu-amd-a8-3500m-ecco-a-voi-llano/31957/1.html*. APU AMD A8-3500M, ecco a voi Llano, 14 giugno 2011.

[9] punto-informatico.it. *http://punto-informatico.it/1586749/PI/News/amd-ati-diventano-una-cosa-sola.aspx*. AMD e ATI diventano una cosa sola, 25 luglio 2006.

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

[11] http://www.nvidia.it/

[12] http://www.amd.com

[13] http://www.hwupgrade.it/

[14] Sergio Congiu *Pàtron editore*. Architettura degli elaboratori, Organizzazione dell'hardware e programmazione in linguaggio assembly, quinta edizione, 2007.

[15] L.Marchetti *http://www.lithium.it/articolo0015p3.htm*. Introduzione alla tecnologia di integrazione SOI, 17 Luglio 2001.