Università degli studi di Padova

Facoltà di Ingegneria

Corso di Laurea in Ingegneria dell'Informazione

A. A. 2012-13

Tesi di laurea triennale

Processori per applicazioni real-time a basso consumo

Real-time and low power consumption application processors

Laureando: Simonato Giosue

Relatore: Bevilacqua Andrea

Indice

1. Sommario

Il documento punta ad analizzare un ramo dell'industria dei semiconduttori che ha occupato negli ultimi anni una fetta sempre crescente del mercato dei dispositivi digitali.

I processori per applicazioni sono microprocessori progettati per svolgere funzioni specifiche, a differenza di un processore *general purpose*. Il loro progetto si basa soprattutto sul raggiungimento di caratteristiche importanti per i sistemi che li utilizzano, caratteristiche che i processori "standard" non hanno necessariamente, e che obbligano i progettisti a compromessi di progettazione.

L'obbiettivo sarà delineare cosa caratterizza un application processor (AP), in base anche ai suoi utilizzi e applicazioni principali. Successivamente verranno analizzati alcuni esempi di chip attualmente in commercio, utili per creare un occhio critico ad un progettista di sistemi digitali.

Infine sarà analizzato il *system on a chip* (SoC) di un sistema embedded di penultima generazione, il quale impiega un processore per applicazioni ARM Cortex come nucleo oltre che moderne tecnologie per raggiungere gli alti standard prestazionali di un dispositivo testa di serie nel mercato di appartenenza. È doveroso fare una precisazione: la penultima generazione a cui ci si sta riferendo, in accordo con la legge di Moore, intende dispositivi annunciati dalla casa costruttrice poco più di un anno fa, dunque ancora pienamente attuali nel mercato mondiale.

2. Panoramica

Applicazioni che elaborano segnali digitali sono onnipresenti nei giorni nostri, la loro diffusione è stata sostenuta da un costante incremento delle prestazioni ed un abbassamento drastico nel tempo dei prezzi.

Un sistema di questo tipo comprende necessariamente una memoria ed interfacce di input ed output, elementi che sono gestiti da una unità di elaborazione centrale detta processore.

Negli ultimi anni siamo stati testimoni di una innovazione sul fronte dell'industria delle comunicazioni e della computazione che ha stravolto il mercato globale: recentemente IDC (*International Data Corporation*) ha riportato che le vendite di *smartphone* hanno raggiunto quota di 144,9 milioni nel primo quarto del 2012 con una crescita del 42,5 per cento rispetto allo stesso periodo dell'anno precedente.

Il numero di smartphone in uso a fine 2012 ha superato il miliardo [1]. Tuttora infatti i prezzi sono pressoché stabili e accessibili da un utente medio, mentre le prestazioni aumentano in modo continuo.

I processori per applicazioni portatili, chiamati anche *mobile* SoC sono il cuore di questa innovazione, e provvedono alla domanda di funzionalità e prestazioni dei dispositivi.

In questo caso il processore viene progettato una ed una sola volta, ovvero è destinato a svolgere le funzioni per cui è stato costruito senza poter esser riprogrammato dall'utente.

Ovviamente in base al tipo di applicazione le caratteristiche del processore varieranno, per esempio nel telefonino intelligente di cui sopra le problematiche relative al tempo di risposta sono preponderanti (sistema real-time).

In questo caso inoltre il processore è integrato nel sistema (*embedded system*) ed è un singolo chip interamente costituito da interconnessioni di più circuiti integrati (pratica preminente al giorno d'oggi). Per sottolineare questo tipo di implementazione si utilizza il termine microprocessore.

In questo capitolo si vuole dare una introduzione su quanto riguarda l'architettura di un generico processore e sul suo modo di ricevere le istruzioni. Dopo di che si passerà alla descrizione di un sistema embedded.

2.1. Il microprocessore

Un microprocessore è attualmente l'implementazione più gettonata della CPU, l'unità centrale dei elaborazione di un computer.

La costruzione dei microprocessori è stata resa possibile grazie all'approccio LSI (*Large Scale Integration*) fondata sulla tecnologia "Silicon Gate Technology": integrare una CPU completa in un solo chip permise di ridurre significativamente i costi dei calcolatori.

Dalla loro introduzione ad oggi, l'evoluzione del microprocessore ha seguito con buona approssimazione la legge di Moore, una legge esponenziale che prevede il raddoppio del numero di transistor integrabili sullo stesso chip ogni 18 mesi.

L'incremento prestazionale verificatosi dalla fine degli anni ottanta però è dovuto soprattutto al miglioramento dell'architettura dei calcolatori attraverso l'adozione di tecnologie RISC (*Reduced Instruction Set Computer)* , all'uso di *pipeline* ed all'implementazione di memorie nascoste all'utente in cui vengono memorizzati dati velocemente reperibili, le *cache*.

La CPU è costituita da due blocchi cardinali che sono quelli che realizzano le funzioni di controllo e quelle aritmetiche e logiche, ed altri due blocchi che consentono l'immagazzinamento dei dati e l'interfacciamento con l'esterno del processore.

- MODULO DI CONTROLLO

  Determina quale particolare azione accade nel processore per un dato istante temporale. Un controllore può essere visto come una macchina a stati finiti; è formato da registri e logica e quindi è un circuito sequenziale. La logica può essere implementata in modi diversi: o come interconnessione di porte logiche fondamentali (usando *standard cell*) o in modo più strutturato usando PLA (*Programmable Logic Array*) e memorie con la sequenza di istruzioni.

- UNITA' DI ELABORAZIONE (ALU)

  Un'unità di elaborazione tipicamente è costituita dall'interconnessione di funzioni logiche combinatorie fondamentali, quali operatori logici (*and, or, xor*) o aritmetici (addizione, moltiplicazione, comparazione, traslazione). I risultati intermedi dell'elaborazione vengono memorizzati in registri.

  Esistono strategie differenti per l'implementazione di unità di elaborazione: celle dedicate e strutturate piuttosto che celle standard automatizzate, o circuiti a connessioni predeterminate

piuttosto che circuiti con interconnessioni flessibili e programmabili. La scelta di una particolare piattaforma implementativa è per lo più influenzata dal compromesso ottimo tra varie figure di merito del sistema, quali area, velocità, potenza, tempo di progetto e grado di riusabilità. L'unita di elaborazione è il nucleo del processore.

- MODULO DI MEMORIA

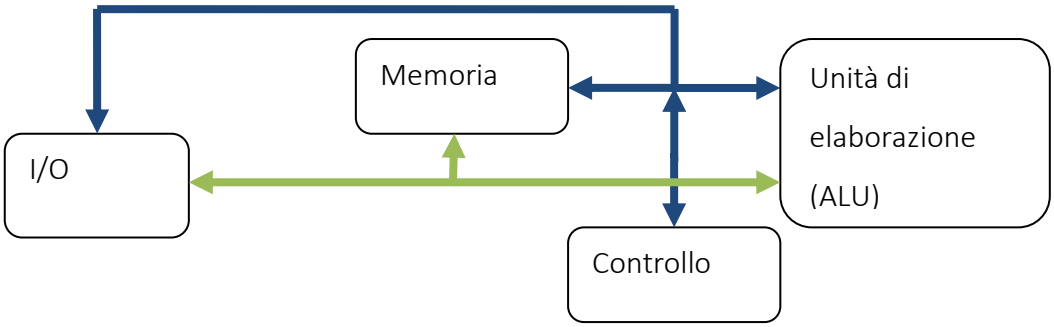
Svolge il ruolo di area centralizzata per l'immagazzinamento dei dati. Esiste una vasta gamma di classi di memorie differenti. La differenza principale tra queste classi risiede nel modo in cui è possibile accedere ai dati, come "a sola lettura" piuttosto che "a lettura scrittura", ad accesso sequenziale piuttosto che casuale, ad accesso a porta singola piuttosto che multipla. Un'altra possibile classificazione delle memorie fa riferimento alla capacità di preservare i dati: per questo proposito le memorie dinamiche devono essere rinfrescate periodicamente , mentre le statiche li conservano finché sono alimentate; infine le memorie non volatili, quali le flash, conservano i dati memorizzati persino quando viene rimossa la tensione di alimentazione. Un singolo processore può sfruttare diverse classi di memoria. Per esempio, le memorie ad accesso casuale possono essere usate per immagazzinare i dati e quelle a sola lettura per memorizzare l'insieme dei comandi.

- RETE DI INTERCONNESSIONI e CIRCUITERIA INGRESSO/USCITA (*I/O*)

Per molto tempo le interconnessioni sono state una preoccupazione dell'ultimo minuto nel flusso di progetto. Sfortunatamente le piste di cui la rete di interconnessione è composta sono sempre meno ideali e presentano carichi capacitivi, resistivi e induttivi alla circuiteria che le pilota.

Con il crescere delle dimensioni del chip, la lunghezza delle piste di connessione cresce di pari passo, causando l'aumento del valore di questi parametri parassiti. Oggigiorno vengono introdotte delle metodologie progettuali automatizzate o strutturate in modo da facilitare lo spiegamento di queste strutture di interconnessione. Alcuni esempi sono i bus integrati su chip, le strutture di interconnessione a maglia e persino intere reti integrate su chip. A livello di schema a blocchi, alcune parti della rete di interconnessioni sono spesso rappresentate in modo astratto (vedi schema che segue), la prima unisce i vari moduli del processore tra loro, mentre la seconda realizza l'interfaccia col mondo esterno.

Tali interconnessioni sono di fondamentale importanza per il funzionamento corretto del sistema, queste comprendono le reti di distribuzione delle alimentazioni e dei clock. Pianificare per tempo queste reti di servizio assicura il funzionamento corretto del circuito, una volta integrato in silicio.



Suddivisione nelle due reti di interconnessione principali in un microprocessore.

2.2. Ciclo di una istruzione

Tipicamente la CPU è l'Interprete del linguaggio macchina. Come tutti gli interpreti, si basa sul seguente ciclo:

- Acquisizione dell'istruzione (*instruction fetch*): il processore preleva l'istruzione dalla memoria, presente nell'indirizzo (tipicamente logico) specificato dal registro PC.

- Decodifica (*operand assembly*): una volta che la word è stata prelevata, viene determinata quale operazione debba essere eseguita e come ottenere gli operandi, in base ad una funzione il cui dominio è costituito dai codici operativi (tipicamente i bit alti delle word) ed il codominio consiste nei brani di microprogramma da eseguire.

- Esecuzione (e*xecute*): viene eseguita la computazione desiderata. Nell'ultimo passo dell'esecuzione viene incrementato il PC: tipicamente di uno se l'istruzione non era un salto condizionale, altrimenti l'incremento dipende dall'istruzione e dall'esito di questa.

La figura sottostante descrive simbolicamente le azioni compiute dalla CPU attraverso blocchi. Si vede chiaramente l'acquisizione dell'istruzione dalla memoria, la decodifica, e successivamente con frecce tratteggiate l'esecuzione vera e propria.

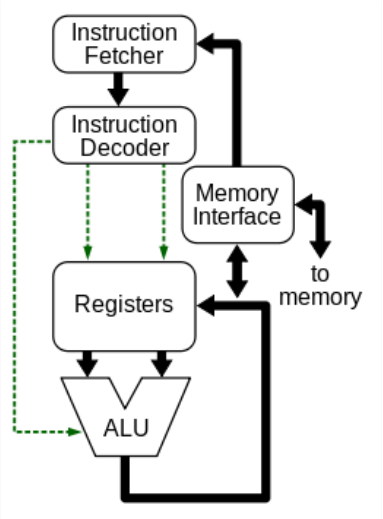


Diagramma a blocchi semplificato di una CPU, dall'alto: il ricevitore di istruzioni, il decodificatore, i registri dati, l'unità aritmetico-logica (ALU); a destra è rappresentata l'interfaccia alla memoria.

Questo ciclo elementare può essere migliorato in vari modi: per esempio, la decodifica di una istruzione può essere fatta contemporaneamente all'esecuzione della precedente e alla lettura dalla memoria della prossima (*instruction prefetch*) e lo stesso può essere fatto con i dati che si prevede saranno necessari alle istruzioni (*data prefetch*).

La stessa esecuzione delle istruzioni può essere suddivisa in passi più semplici, da eseguire in stadi successivi, organizzando la unità di controllo e la ALU in stadi consecutivi, come delle catene di montaggio (pipeline): in questo modo più istruzioni possono essere eseguite quasi contemporaneamente, ciascuna occupando ad un certo istante uno stadio diverso della pipeline.

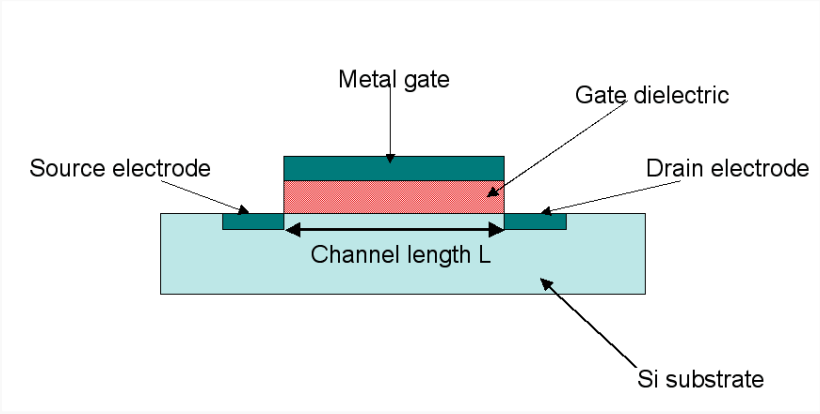
Il problema di questo approccio sono le istruzioni di salto condizionato: la CPU non può sapere a priori se dovrà eseguire o no il salto prima di aver eseguito quelle precedenti, così deve decidere se impostare la pipeline tenendo conto del salto o no: e in caso di previsione errata la pipeline va svuotata completamente e le istruzioni in corso di decodifica rilette da capo, perdendo un numero di cicli di clock direttamente proporzionale al numero di stadi della pipeline. Per evitare questo i processori moderni hanno unità interne (*"Branch prediction unit"*) il cui scopo è tentare di prevedere se, data una istruzione di salto condizionato e quelle eseguite in precedenza, il salto dovrà essere eseguito o no.

Inoltre i processori possono implementare al loro interno più unità di esecuzione per eseguire più operazioni contemporaneamente (architetture dualcore, multicore, quadcore, sixcore). Questo approccio incrementa le prestazioni delle CPU ma ne complica notevolmente l'esecuzione, dato che per poter eseguire in modo efficiente più operazioni in parallelo la CPU deve poter organizzare le istruzioni in modo diverso da come sono organizzate dal programmatore. In un sistema multiprocessore tutti i processi che girano sulle varie CPU condividono un unico spazio di indirizzamento logico, mappato su una memoria fisica che può però essere distribuita fra i vari processori. Ogni processo può leggere e scrivere un dato in memoria semplicemente usando un comando *load* o *store*, e la comunicazione fra i processi avviene attraverso la memoria condivisa.

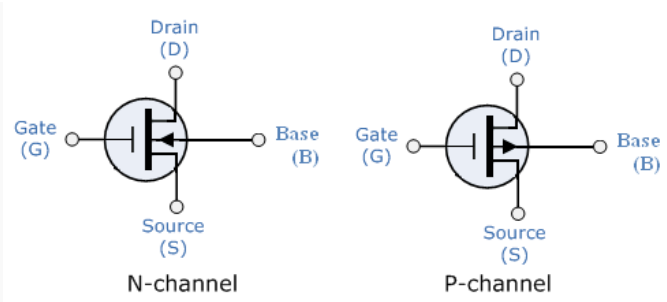
2.3. Architettura

L'integrazione di larga scala ha potuto ridurre non di poco i costi di produzione, ed ha consentito quindi la distribuzione di massa dei dispositivi digitali a cui siamo tanto abituati.

I microprocessori sono costruiti con transistor MOSFET (*metal oxide semiconductor field effect transistor*). Nell'immagine sottostante si può vedere come è costituito un MOSFET, questi vengono impressi attraverso il processo di fotolitografia che consiste nel lavorare una fetta di silicio monocristallino di spessore inferiore al millimetro. La fotolitografia impiega maschere sub-micrometriche in maniera ciclica per selezionare le parti da, rispettivamente: drogare, incidere, ossidare. Il processo di produzione ad oggi ha raggiunto le poche decine di nanometri di dimensioni per transistor ed il loro numero ha superato il milione di elementi per circuito integrato.



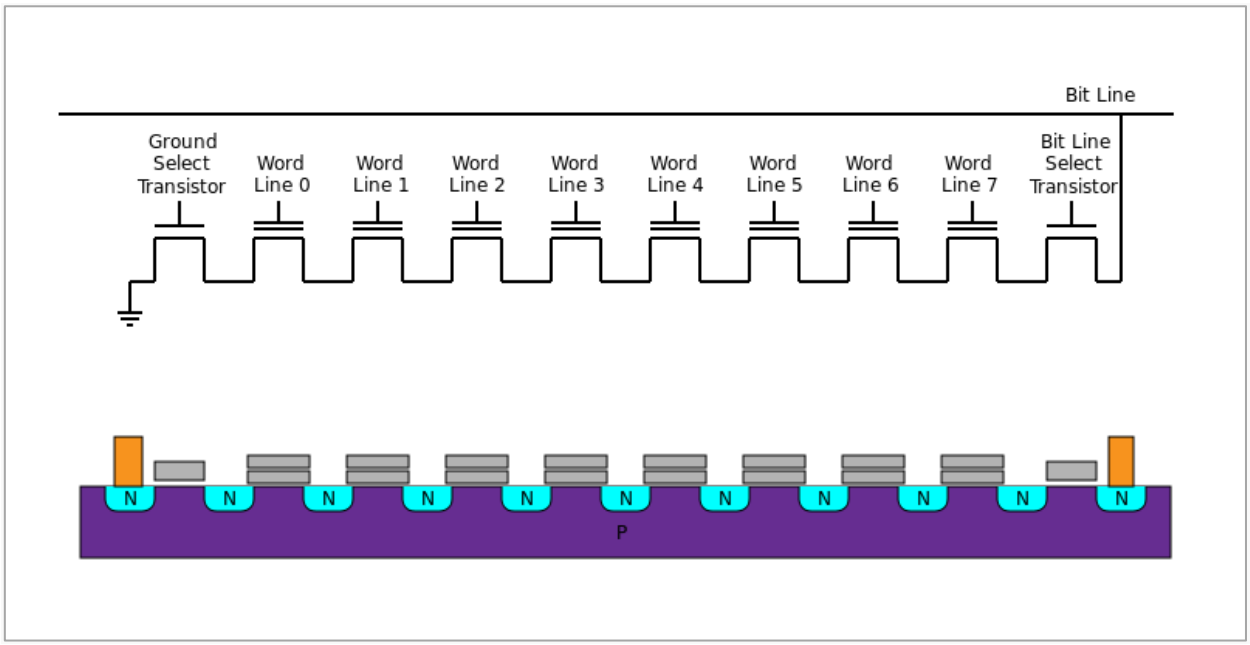
Sezione trasversale di un generico transistor MOSFET cresciuto su di una fetta di silicio dello spessore di 0,5 mm.



Simbolo circuitale di un MOSFET di tipo n (ovvero cresciuto in un substrato p-drogato, con elettrodi n-drogati), e simbolo di un MOSFET di tipo p (con caratteristiche complementari al precedente).

Tali transistor costituiscono ogni singola parte di un circuito integrato; per esempio registri, ALU, memorie flash addirittura integrate all'interno di uno stesso microprocessore e molti altri circuiti

digitali. Logiche diverse possono venire utilizzate per implementare il nostro sistema digitale. Vale la pena citare la logica MOS complementare (CMOS), la logica a rapporto e la logica dinamica. Un esempio si può vedere nell'illustrazione successiva: una generica bitline di una memoria non volatile.



Esempio di flash a NAND che utilizza transistor MOSFET. I transistor disegnati con una ulteriore linea tra gate e substrato indicano transistor di tipo n a gate flottante. Questi consentono di conservare l'informazione anche una volta tolta l'alimentazione.

La densità di integrazione e le prestazioni dei circuiti integrati hanno avuto una crescita stupefacente negli ultimi decenni, la capacità delle memorie è aumentata di 1000 volte dal 1970 ad oggi, come la già citata legge di Moore aveva predetto. Così come il numero di transistor per circuito integrato che ha sorpassato il milione di elementi.

In base all'uso richiesto e al tipo di segnali di ingresso dobbiamo basare la scelta del processore da impiegare. I processori per applicazioni vengono ampiamente utilizzati nei sistemi embedded.

La più importante casa costruttrice di APs è ARM Holdings, attualmente essa ricopre il 75% del mercato mondiale dei processori a 32 bit [2].

L'architettura ARM utilizza l'approccio di tipo RISC. È un set di istruzioni ridotto per eseguire operazioni semplici in tempi brevi. Questo approccio di programmazione fa sì che vengano messe da parte istruzioni più complesse e vengano sostituite da istruzioni molto più semplici e basilari. In una macchina di questo tipo le uniche operazioni che permettono di accedere alla memoria sono quelle di *load* e di *store*, tutte le altre utilizzano registri. Un codice di questo tipo permette di

realizzare programmi molto veloci e soprattutto ottimizzati, a scapito di una più pesantezza nella sintassi.

Vengono inseriti nell'*instruction set* del microprocessore istruzioni anche molto complesse per simulare le funzioni di alto livello dei linguaggi di programmazione direttamente nei processori.

L'architettura RISC ha dalla sua parte la velocità di esecuzione del programma ma lo svantaggio è l'occupazione di memoria da parte del codice, pur avendo integrato istruzioni complesse all'interno della CPU.

Un altro tipo di architettura si basa sulla tecnologia CISC (*complex instruction set computer*) dove le istruzioni sono complicate in modo da realizzare microprocessori che compiano istruzioni paragonabili ai linguaggi di alto livello. La realizzazione pratica richiede molto più silicio rispetto all'architettura RISC, ma rispetto a queste diminuisce la difficoltà di progetto del compilatore e del sistema operativo, in quanto il microprocessore utilizza istruzioni complesse (pseudo-istruzioni).

2.4. Sistema embedded

Un sistema embedded come dice la parola stessa è un sistema inglobato. Si intendono quei sistemi elettronici a microprocessore progettati appositamente per una determinata applicazione. Dunque un sistema che usa al suo interno un application processor è un sistema embedded.

Questo tipo di sistemi come si intuisce hanno dei compiti noti già durante lo sviluppo, che verrà eseguito quindi in modo da garantire una combinazione hardware/software specificamente studiata per la tale applicazione. Questo consente di ridurre considerevolmente l'hardware connesso alla scheda e anche il costo di fabbricazione.

Infatti essere un processore per applicazioni abbastanza limitate, rende questi molto economici e molto più semplici da inserire all'interno di un circuito elettrico rispetto ai sistemi general purpose quali i microprocessori di un computer.

L'esecuzione del software avviene in tempo reale per permettere un controllo deterministico dei tempi di esecuzione e questo influisce sulle prestazioni minime che può avere un questo tipo di sistema.

La progettazione dei sistemi dedicati dipende anche dalla loro distribuzione alla collettività. Se la tiratura è limitata, come nel caso di macchinari specifici e costosissimi come apparati elettromedicali che si possono permettere solo i migliori ospedali, questi sistemi sono realizzati con l'impiego del miglior hardware presente sul mercato; rinunciando al risparmio e garantendo una qualità senza confronti (in altre parole il costo influisce molto sulla progettazione e costruzione). Per tirature elevate come cellulari e walkman i costi di produzione verranno ammortizzati dal numero di copie che si riesce a vendere, e in tal caso si eseguirà un attenta scelta dell'hardware dopo un approfondita analisi della richiesta che offre il mercato.

Per la grande maggioranza dei sistemi embedded non si parla di software, bensì di *firmware* (*firm* stabile, *ware* componente), che trova solitamente posto in una memoria non volatile o in una ROM.

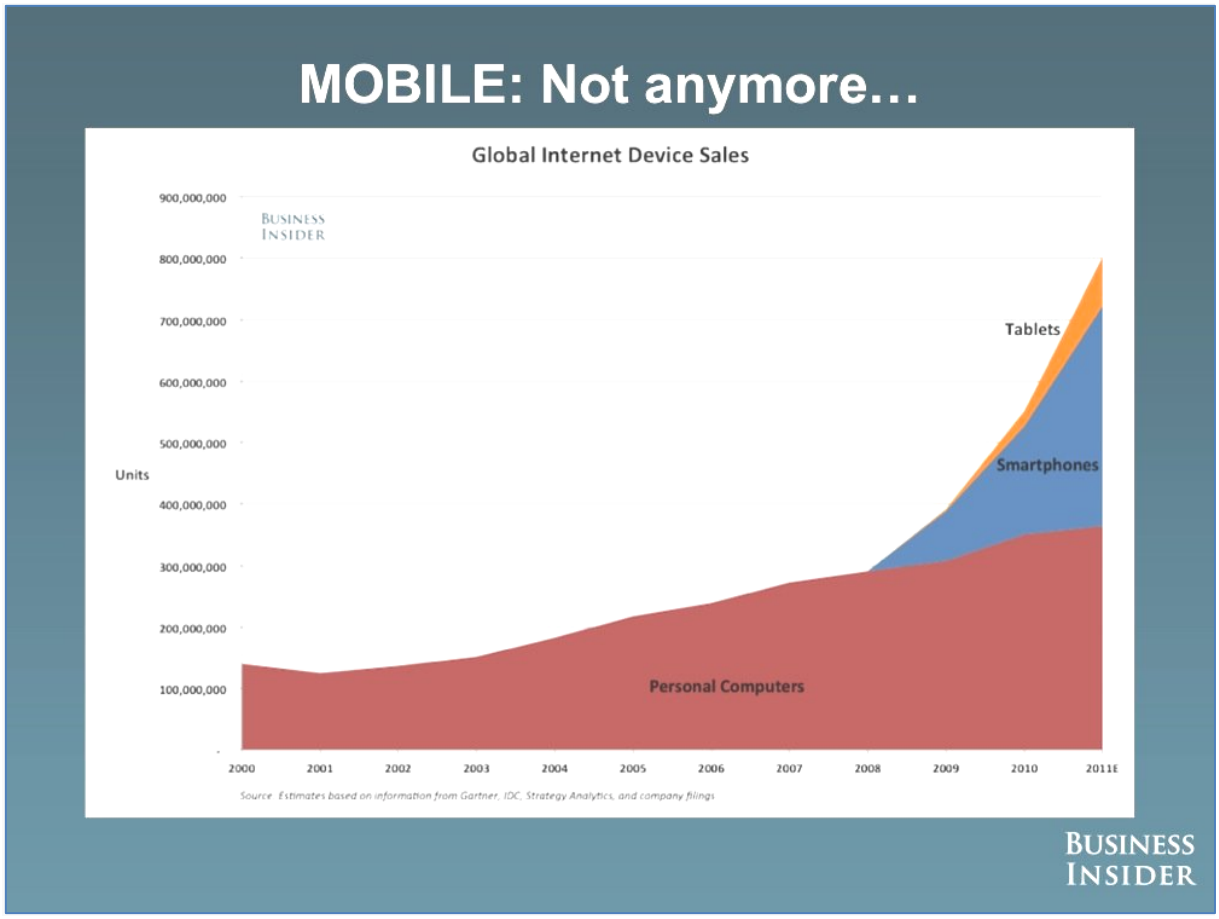
Per i sistemi embedded a grandi volumi di produzione ci si sta spostando verso i cosiddetti SoC (*System on a Chip*). I SoC racchiudono in un singolo circuito integrato tutte le periferiche e la CPU stessa. Se non può esser realizzato un SoC un alternativa risiede nei SiP (*System in Package*), ovvero un singolo package che racchiude in sé diversi circuiti integrati.

I più importanti processi di fabbricazione di microprocessori per dispositivi portatili utilizzano proprio la filosofia SoC: accoppiano il processore centrale con circuiti specializzati per la comunicazione, la grafica, navigazione, e altri strumenti. I sistemi SoC permettono di risparmiare

energia perché sono costituiti da sistemi dedicati imballati in stretta collaborazione. La miniaturizzazione standard di un processore per PC non garantirebbe lo stesso risparmio di energia [3].

Un esempio è Atom chip di Intel che appena uscito (aprile/maggio 2011, processo di 45 nm) non era in grado di raggiungere le prestazioni in termini energetici dei chipset di casa Qualcomm (basati su processori ARM). Il budget in potenza tipico di uno smartphone è 1 watt; Atom originale ne consumava tanto anche in uno stato inattivo (di *stand-by*). La casa di produzione Intel nel 2011 ha prodotto solo 1 milione dei 760 milioni di chip venduti per l'uso in dispositivi portatili.

Da tener conto pero è il fatto che una potenza mondiale produttrice di processori per computer e server come Intel, stia cercando di immettersi nel mercato dei dispositivi portatili. Il motivo è che negli ultimi anni le vendite dei PC si sono stabilizzate mentre sono aumentate drasticamente quelle dei dispositivi che implementano sistemi embedded, come si può notare dal grafico.



Uno sguardo al mercato dei dispositivi elettronici più richiesti degli ultimi anni. [4]

3. Aspetti peculiari di un AP

Le caratteristiche che deve avere un generico AP dipendono soprattutto dall'utilizzo per cui è stato costruito, ovvero per quale applicazione specifica questo tipo di processore è nato.
Il contesto in vengono utilizzati è davvero molto vasto, ma tra tutte le implementazioni spiccano sicuramente i dispositivi cellulari più recenti e i sistemi *touch screen* portatili quali i tablet.
Un altro tipo di sistema che si affida a questa tipologia di processori sono per esempio gli impianti elettronici delle automobili, presenti ormai in quasi ogni tipo di vettura. Ciò che controlla il display elettronico che interagisce con l'utente e nel frattempo carica la musica della memoria flash collegata nell'apposita presa USB, è un microprocessore con un dato firmware. Un secondo tipo tipo di application processor usato nelle auto invece è quello che comanda l'unità di controllo elettronico presente in tutti i moderni autoveicoli, le sue funzioni sono assistere i sistemi di stabilità del veicolo, ABS, air bag, anti collisione (o sensori di parcheggio).
Ovviamente la capacità di evitare un guasto che comprometta il corretto funzionamento è molto più importante nella seconda tipologia di AP per auto, il processore deve cioè avere maggiore affidabilità (*system reliability*).
Riguardo al firmware se confrontato con quello di un recente smartphone, questo sarà molto diverso: quello del dispositivo mobile infatti sarà molto più complesso dato che deve organizzare le chiamate e il sistema Wi-Fi; mentre l'utilizzatore sta scattando foto o giocando con display ad alta definizione. In questo caso deve quindi garantire prestazioni maggiori oltre che consumi ridotti della batteria. Il microprocessore ricopre un ruolo centrale: è il cuore del nostro dispositivo.
Altri utilizzi notevoli di un AP possono essere una fotocopiatrice o un modellino aereo, in entrambi i casi esso svolge i compiti per i quali è stato progettato: coordinare i vari programmi possibili di stampa o controllare motore, ricevere il segnale wireless del telecomando e azionare la virata di conseguenza. I sistemi sono entrambi embedded ma nel caso della stampante il problema alimentazione non sussiste o comunque è relativo, ciò che è importante è una buona interfaccia per interagire con l'utente. Nell'aereo-modellino la questione è invertita: non è necessaria un interfaccia con l'utente sul velivolo, bensì che il sistema sia portabile per il tempo assicurato.
Nel secondo caso siamo di fronte ad un sistema profondamente integrato real-time che risponde ai comandi in tempi più brevi possibile, mentre nel primo invece la rapidità di risposta non è preponderante, anche se non possiamo trascurare la pazienza dell'utente che usa la fotocopiatrice.

Comunque se il prezzo del SoC dovesse dipendere solo dalla velocità di risposta, la scelta in questo caso sarebbe più rilassata.

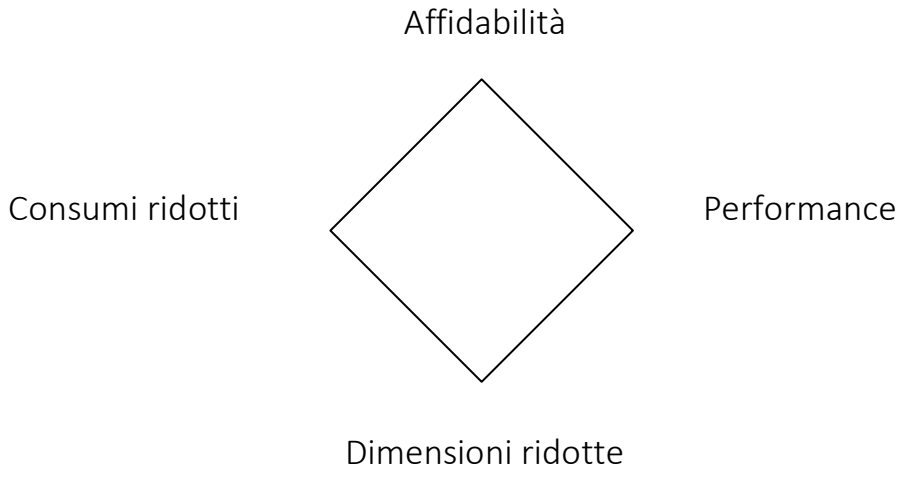
Alcuni application processors sono molto più vicini a somigliare ad un processore general purpose di altri, questo è vero per esempio, per i *mobile* AP degli smartphone e dei tablet di ultima generazione.

Sebbene il sistema operativo sia pressoché immutabile una volta deciso e non sostituibile con facilità (per un utente medio questo è sicuramente vero), non solo è soggetto ad aggiornamenti ma può anche installare le applicazioni più svariate.

Le differenze sostanziali tra mobile APs e processori standard stanno nell'importanza di contenere la potenza dissipata, nelle dimensioni, nelle temperature raggiungibili dal sistema, nella frequenza di utilizzo.

Le prestazioni, la velocità di esecuzione e dunque la frequenza del microprocessore di uno smartphone sono infatti limitate dall'utilizzo di energia, che è vincolata alla capacita della batteria. Un dispositivo portatile con capacità computazionali paragonabili a quelle di un processore Intel i7 (orientato ai personal computer con 2,4 GHz di frequenza base) non avrebbe mercato con la batteria che possiede: la sua portabilità sarebbe limitata a qualche ora. La soluzione sarebbe aumentare la dimensione della batteria, ma questo porterebbe ad avere un cellulare troppo pesante o troppo grande.

Ipotizziamo di avere un quadrato ed una area massima possibile per lui. Ad ogni angolo inseriamo rispettivamente: Affidabilità, Consumi ridotti, Performance, Dimensioni ridotte; non in maniera casuale ma in modo che ad angoli opposti ci siano caratteristiche incompatibili, le quali necessitano di un compromesso.



Supponiamo di tirare l'angolo delle prestazioni, i consumi aumenteranno perché l'area del quadrato deve rimanere la stessa. Allo stesso modo diminuire le dimensioni porta ad avere meno controlli sulla produzione del dispositivo e quindi ad aumentare la probabilità di errori dunque renderlo meno affidabile. L'area del quadrato è la tecnologia di cui disponiamo nel progetto di un processore per applicazioni.

3.1. System reliability

"Reliability is, after all, engineering in its most practical form"

James R. Schlesinger

Mentre un eventuale guasto è influenzato da parametri stocastici, la qualità, l'affidabilità e sicurezza non sono raggiunti da matematica e statistica, l'approccio è meno diretto ed utilizza studi e test associati all'analisi dei dati raccolti. Per un sistema complesso si utilizzano strumenti hardware-software specifici, l'analisi dei fallimenti, la documentazione tecnica e l'esperienza umana.
Rendere le unità di controllo elettronico delle auto affidabili e forti ai guasti è un obbiettivo preponderante per i progettisti di questi sistemi. L'affidabilità di un sistema descrive la capacità di funzionare per un dato periodo sotto condizioni standard, eventualmente considerando brevi periodi di condizioni straordinarie .
Per creare un processore con bassa probabilità al guasto è molto importante conoscere bene le funzioni che dovrà svolgere l'application processor ma soprattutto l'ambiente in cui si appresta a lavorare. Per esempio la temperatura di lavoro in cui deve garantire il funzionamento è diversa in un contesto automobilistico o in un contesto indirizzato direttamente al cliente come un dispositivo portatile. Mentre nel primo caso ci si appresta a garantire l'affidabilità del componente in un range di temperatura tra -10 e 150 gradi centigradi, nel secondo caso da 0 a 40 è sufficiente. I costi seguiranno le capacità del componente in quanto verranno utilizzate tecnologie e materiali più all'avanguardia per garantire più alti standard.
Ci sono dei passi che è consuetudine seguire per iniziare il progetto di un microprocessore affidabile:

- Descrizione dei meccanismi di fallimento conosciuti attraverso modelli matematici ricavati dall'analisi a priori;

- Determinazione del profilo di carico degli errori riguardo all'applicazione specifica, nei test;

- Simulazione dell'affidabilità del componente utilizzando i modelli e il profilo di carico nei test precedentemente trovati.

Gli errori possono essere divisi in *soft errors* e in *hard errors*.

- SOFT ERRORS: errori di transizione o di singolo evento. Sono errori nel processo di esecuzione dovuti a rumore elettrico o radiazioni esterne, piuttosto che dovuti al progetto o al processo di fabbricazione. Sebbene possano causare sbagli nella computazione e nella corruzione dei dati, questi errori non danneggiano il microprocessore e non sono visti come una grossa preoccupazione sull'affidabilità nella vita del dispositivo.

- HARD ERRORS: causati da difetti nella metallizzazione o nella siliciurizzazione del microprocessore. Sono errori permanenti una volta manifestati. A loro volta si possono suddividere in errori estrinseci dovuti al processo di costruzione ed errori intrinseci dovuti alla vita del dispositivo.
  Un esempio d'errore è la contaminazione nel reticolo cristallino della superficie di silicio o imperfezioni su questa che possono provocare la rottura dell'ossido tra gate e substrato in un transistor MOSFET con conseguente percorso conduttivo tra i due.

Un modo per scovare gli errori è una tecnica chiamata *burn-in*. Il processore viene testato a temperature e tensioni elevate in modo da velocizzare la manifestazione di errori estrinseci. Così facendo i dispositivi semiconduttori in vendita risultano avere un alto tasso di esenzione da questo tipo d'errore. È importante altresì saper descrivere i possibili errori intrinseci, essi infatti riducono la vita del dispositivo; una buona percentuale è dovuta allo *scaling* delle dimensioni: la riduzione per esempio dello spessore nell'ossido di gate consente a impurità meno profonde di creare cortocircuiti indesiderati. Allo stesso modo un'elevata densità d'integrazione aumenta il numero di transistor che possono contenere errori e riduce la capacità di rilevarli.
Tecniche utilizzate per la rivelazione di errori intrinseci sono *Reliability Aware MicroProcessor* (RAMP) sviluppata in casa IBM, e *Dynamic Reliability Management* (DRM) [6].
La questione sull'affidabilità interessa molto più gli application processor rispetto i general purpose in quanto sia nel caso di un dispositivo mobile che nel caso di applicazioni specifiche (aereo-modellino discusso in precedenza per esempio) è più probabile essere in condizioni extra-ordinarie.
Importante è la gestione di un aspetto rilevante quale la temperatura degli APs. In questo caso infatti molto spesso la condizione di lavoro non consente una dissipazione ottimale del calore, in un

sistema embedded lo spazio è necessariamente limitato ed un dissipatore come quelli presenti in un personal computer potrebbe essere troppo grande.

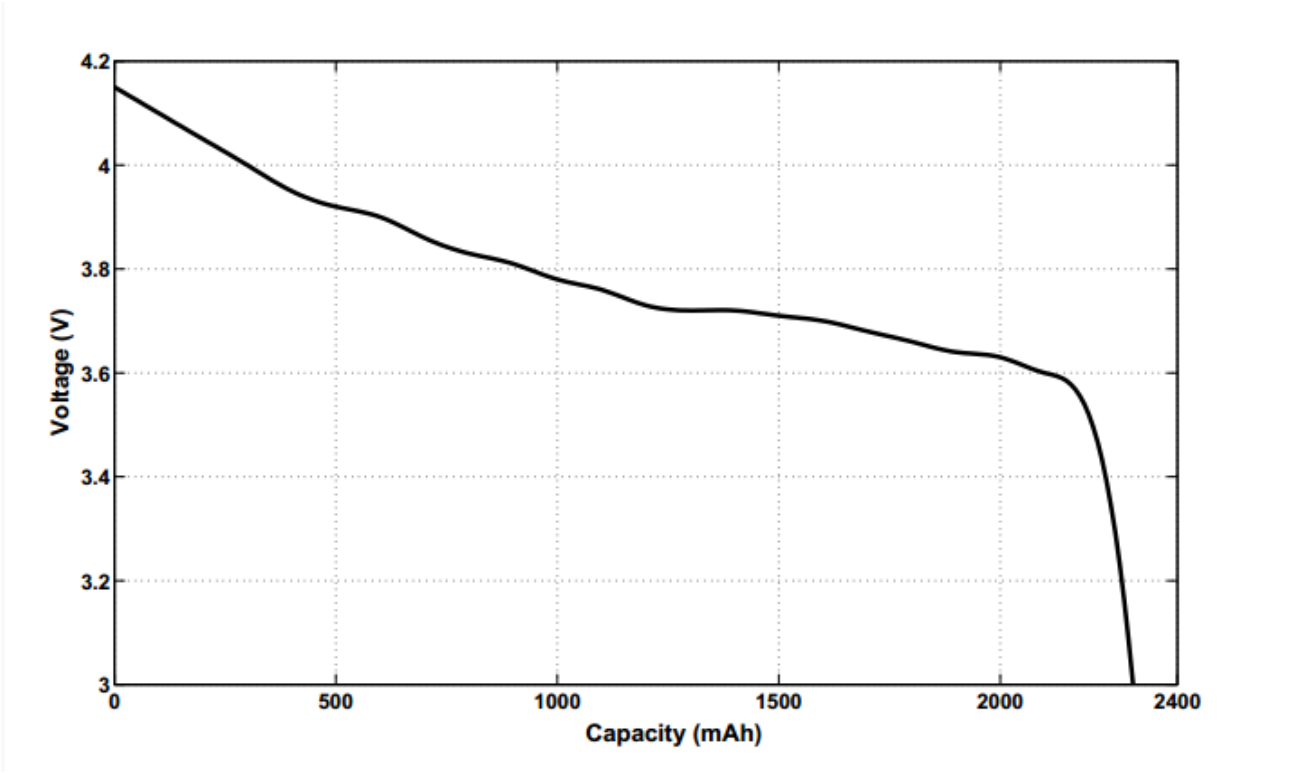
Una soluzione è l'unità di controllo della temperatura (TMU) presente nei più moderni dispositivi mobile che limita la crescita della temperatura se questa raggiunge una certa soglia critica, abbassando le prestazioni. L'analisi di un *Thermal Management Unit* in uso su di un microprocessore di un recente smartphone verrà affrontata nel capitolo 5.

3.2. Power delivery

"Quando non c'è energia, non c'è colore, non c'è forma, non c'è vita."

Michelangelo Merisi da Caravaggio

L'efficienza energetica dei circuiti integrati continua ad essere un fattore determinante nella scelta delle dimensioni, del peso e dei costi di sistemi elettronici portatili. Questi vengono alimentati con batterie al litio ricaricabili, tramite intermediazione di convertitori DC-DC che mantengono la tensione al livello voluto. Infatti la tensione in uscita nelle batterie al litio varia in base alla carica ancora immagazzinata, come si può vedere nel grafico sottostante.



Tipica scarica di una batteria al litio.

Molte delle funzionalità dei cellulari di ultima generazione usano diversi circuiti e blocchi funzionali ognuno dei quali è alimentato con una specifica tensione.

Nell'evoluzione dei dispositivi mobile l'elaborazione dei dati in formato digitale ha preso un ruolo sempre crescente nella frazione dei consumi. Per esempio nella seconda generazione di *code division multiple access* (CDMA) ovvero il 2G, la banda base digitale e il circuito di memoria occupava circa il 10 per cento della potenza totale. Nella terza generazione (3G) questa percentuale

è salita al 30-50 per cento dei consumi totali di potenza, dato che le funzioni associate a filtri o flusso dei dati sono adesso attuate con circuiti digitali.
Ridurre il consumo di potenza della banda digitale e del circuito di memoria è di fondamentale importanza per aumentare la portabilità in senso temporale del dispositivo ed aumentare la vita della batteria.
Scalare la tensione di alimentazione è il modo più semplice per ridurre anche i consumi nei circuiti integrati. Specialmente in un circuito digitale l'energia attiva $E_{ATT}=CV^2$ richiesta per completare un'operazione si riduce quadraticamente con la tensione di alimentazione. Allo stesso tempo l'abbassamento di questa comporta una riduzione nei tempi di operazione.
Un altro metodo è il cosiddetto *power gating* che consiste nell'abbassare la frequenza o spegnere la corrente temporaneamente nei blocchi del sistema che non sono utilizzati. Il power gating viene organizzato dall'unità di controllo della potenza, un microcontrollore a se stante con un proprio firmware. Le sue funzioni principali sono:

- Monitorare le connessioni di alimentazione carica della batteria;
- Controllare la potenza dei circuiti integrati;
- Spegnere eventuali componenti non necessarie nel sistema, quando queste sono inattive;
- Controllare le funzioni di stand-by del dispositivo;
- Regolare il clock.

Un esempio del suo funzionamento sarà analizzato nel capitolo 5.

4. Il mercato dei processori per applicazioni

Molto spesso grandi case costruttrici di dispositivi digitali, quali palmari o tablet, non costruiscono da se il cuore del dispositivo, ma si affidano a specialisti. Per far questo è necessario avere un'idea di cosa il mercato ci offre per il nucleo del sistema che si vuole progettare. Analizziamo quindi alcuni esempi dei più recenti processori per applicazioni.

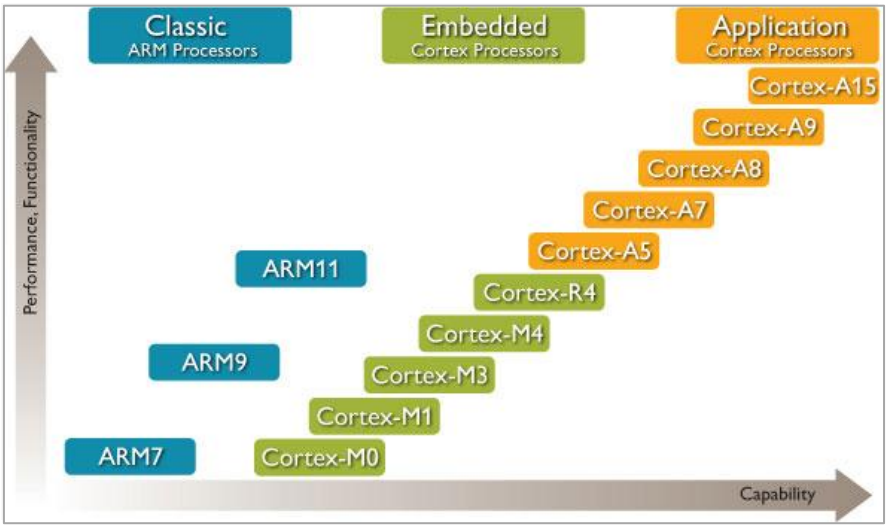
4.1. ARM Cortex

Iniziamo da alcuni processori a disposizione di casa ARM Holdings.

Storicamente l'acronimo ARM significava Acorn RISC Machine, modificato successivamente in Advanced RISC Machine e poi in ARM Holdings. Come già anticipato è una famiglia di microprocessori RISC a 32 bit, utilizzata prevalentemente nei sistemi embedded, dove i bassi consumi sono all'ordine del giorno.

L'ARM Cortex è una famiglia di microprocessori presentata nel 2005 formata da una serie di blocchi funzionali che possono essere collegati tra loro al fine di soddisfare le esigenze dei clienti. I processori Cortex sono disponibili in configurazione singolo core o multicore e per ogni famiglia esistono più core con prestazioni diverse.

La famiglia Cortex è suddivisa nella serie A (*Application*), la serie R (*Realtime*) e la serie M (*Microcontroller*), e come si può notare nell'immagine la scelta è abbastanza ampia.



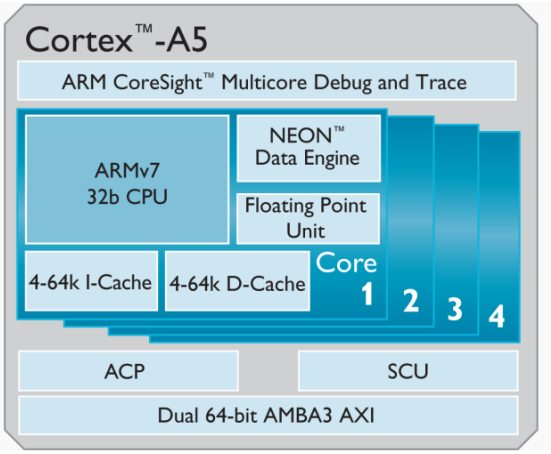
Prestazioni/capacità di alcuni processori ARM. [8]

4.1.1. ARM Cortex A-series

Questo gruppo è formato dai microprocessori indirizzati a telefoni cellulari evoluti come gli smartphone ma anche tablet e applicazioni che necessitano di potenza di calcolo e flessibilità.

- ARM CORTEX-A5

  È il processore più piccolo, con costi minori e un basso consumo; ha la capacità di fornire internet ad un ampia gamma di dispositivi mobili anche di fascia medio-bassa (Samsung Galaxy Ace, HTC Desire, Huawei Ascend, LG Optimus, Sony Experia).



Struttura processore Cortex-A5.

Organizzato in 1-4 core basato sul set di istruzioni ARMv7, tramite l'unità NEON esegue la stessa operazioni in più dati contemporaneamente (SIMD, *single processor multiple data*). Inoltre l'architettura Floating Point fornisce supporto hardware per operazioni *floating point* nelle tre varianti di precisione (*half-, single-, double-*). Supporta anche operazioni di tipo Thumb e Thumb-2. Ha una pipeline a 8 stadi.

ARM CoreSight Multicore Debug and Trace fornisce il supporto necessario per il tracciamento di queste istruzioni mentre AMBA3 AXI consente l'interfacciamento e la comunicazione con le periferiche.

La serie di processori Cortex-A5 è stata presentata in due versioni di chip: la TSMC 40G, indirizzata alla produzione di architetture complesse, votate alle prestazioni; mentre il TSMC 40LP è indirizzato ai chip che fanno nel risparmio energetico il loro punto di forza.

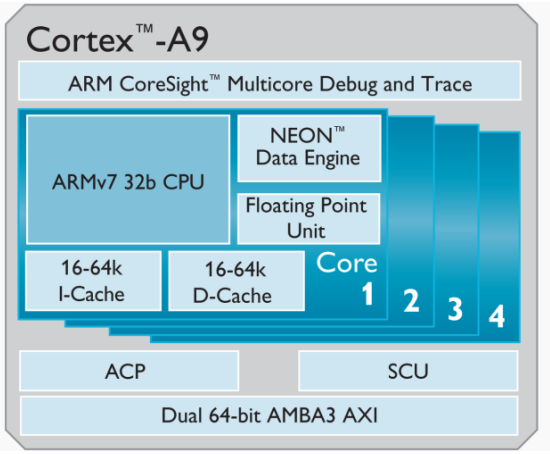
In dettaglio le caratteristiche principali nella tabella che segue.

ARM Cortex-A5 Prestazioni in potenza ed area		
	TSMC 40LP	**TSMC 40G**
Tipo di processore/Tensione nominale	Bassi consumi / 1,1V	Prestazioni / 1,0V
Ottimizzazione in frequenza o nelle prestazioni	Frequenza	Frequenza
Frequenza	530-600 MHz	>1GHz
Area senza RAMs/cache	0,27 mm^2	0,27 mm^2
Area con 16K/16K cache	0,53 mm^2	0,53 mm^2
Area con 16K/16K cache + NEON	0,68 mm^2	0,68 mm^2
Potenza dinamica	0,12 mW/MHz	<0,08mW/MHz
Efficienza energetica	13 DMIPS/mW	>20 DMIPS/mW

Parametri ARM Cortex-A5 nelle due varianti proposte.

- ARM CORTEX-A9

  Un altro componente della famiglia Cortex A-series indirizzato a scelte di efficienza energetica e alte prestazioni di dispositivi di fascia medio-alta è il Cortex-A9.



Struttura processore Cortex-A9.

È un processore multicore che può avere fino a 4 core Cortex basati su istruction set ARMv7 e dotati di gestione coerente della cache.

Istruzioni SIMD NEON sono in grado di eseguire fino a 16 operazioni per istruzione, mentre il set di istruzioni Thumb-2 riduce la dimensione dei programmi con una minima riduzione delle prestazioni. Ha due pipeline a 8 stadi.

Un esempio concreto dell'utilizzo del processore Cortex-A9 è Apple A5, SoC progettato da Apple Inc. e prodotto da Samsung. Apple A5 è il processore utilizzato nel iPhone 4S, ed anche nell'iPad Mini.

Il contenitore dell'A5 racchiude 512 MB di memoria a basso consumo DDR2 a 533 MHz ed alcune unità come la ISP (*image processing*) utilizzata per operazioni di processione delle immagini e l'unità earSmart.

Inoltre viene integrato nell'implementazione del sistema embedded Exynos, che verrà analizzato in dettaglio nel prossimo capitolo. Exynos è l'unita centrale con cui sono stati costruiti gli smartphone di penultima generazione Samsung Galaxy SII e Samsung Galaxy SIII. La differenza tra i due risiede soprattutto nel processo di fabbricazione del rispettivo SoC Exynos: con un passo di 45 e di 32 nm rispettivamente.

Il processore ARM Cortex-A9 consuma meno di 250 mW per core alla frequenza di 1GHz.

Anche in questo caso ARM Holdings ha proposto due implementazioni del suo prodotto. Il processore viene prodotto utilizzando attualmente un processo a 32 nm. Si noti come la frequenza di utilizzo influisca molto sulle prestazioni cosi come sui consumi.

ARM Cortex-A9 Prestazioni in potenza ed area			
	Cortex-A9 Single Core Soft Macro Trial Implementation	**Cortex-A9 Dual Core Hard Macro Implementations**	
Processo	TSMC 65G	TSMC 40G	
Ottimizzazione	Ottimizzazione delle Prestazioni	Prestazioni	Potenza
Libreria Standard Cell	ARM SC12	ARM SC12 + High Performance Kit	ARM SC12 + High Performance Kit
Prestazioni	2075 DMIPS	10000 DMIPS	4000 DMIPS
Frequenza	830 MHz	2000 MHz (tipicamente)	800 MHz (tipicamente)
Efficienza energetica (DMIPS / mW)	5,2	5,26	8,0
Potenza totale alla frequenza ottimale	0,4 W	1,9 W	0,5 W
Area di silicio	1,5 mm^2 (esclusa la cache)	6,7 mm^2	4,6 mm^2

Parametri ARM Cortex-A9 nelle configurazioni proposte.

- ARM CORTEX-A12 e ARM CORTEX-A15

  Vale la pena accennare a questi processori perché nel prossimo futuro potrebbero essere il nucleo dello smartphone più diffuso. Secondo ARM infatti il nuovo Cortex-A12 sarà presente su 580 milioni di dispositivi entro il 2015. La nuova architettura Cortex-A12 è ottimizzata per le tecnologie di processo a 28 nm e può essere implementata fino a soluzioni quad-core.
  Esso è destinato a prendere il posto di ARM Cortex-A9 .
  Per quanto riguarda Cortex-A15 questo è già stato utilizzato per il Samsung Galaxy S4, uscito nella primavera del 2013 e successore del Samsung Galaxy SIII.

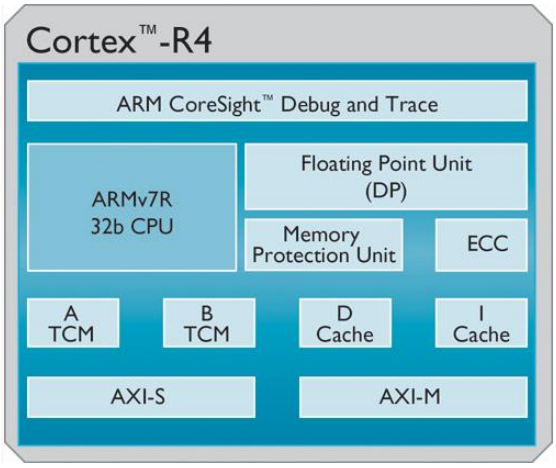
4.1.2. ARM Cortex R-series

La serie R di processori in tempo reale ARM offre soluzioni di calcolo ad alte prestazioni per sistemi embedded, dove l'affidabilità, l'alta disponibilità, la tolleranza ai guasti, la manutenibilità (capacità del sistema di essere facilmente ripristinato qualora sia necessario realizzare un intervento di manutenzione) e risposte in tempo reale sono obbligatori.

- ARM CORTEX-R4

  Il processore Cortex-R4 è il primo processore profondamente embedded real-time basato sull'architettura ARMv7-R cioè sviluppato per sistemi incorporati e applicazioni in tempo reale. Questo processore è ampiamente utilizzato dall'azienda statunitense Broadcom Corporation operante nel settore dei semiconduttori, nei circuiti integrati e nelle reti di telecomunicazioni. Non è inoltre l'unica azienda che si affida a implementare sistemi usando la serie Cortex-R in generale, ne troviamo almeno un'altra ben nota come la Texas Instruments.

  Applicazioni possono essere le più svariate: delle unità di controllo degli airbag nelle automobili, ai lettori Blu Ray e lettori mp3 fino a supporti per hard disk.



Struttura processore Cortex-R4.

Le novità rispetto alle configurazioni Cortex-A sono un interfaccia TMC (*Tightly-Coupled Memory*) utile per reperire dati da memorie esterne o scrivere dati su queste memorie. L'interfaccia ATCM contiene il codice per la gestione delle interruzioni o delle eccezioni che possono essere sollevate in seguito ad un errore. L'interfaccia BTMC invece contiene i

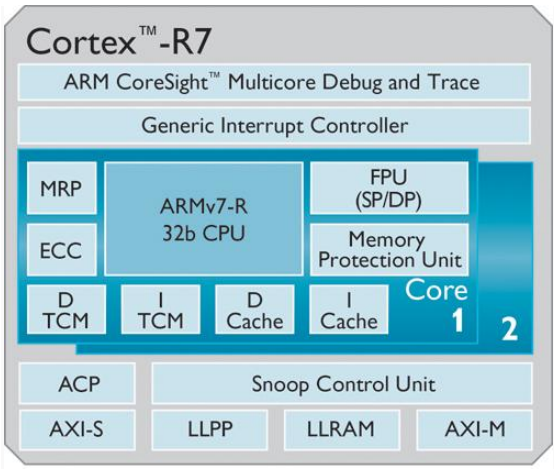
blocchi di dati che devono essere elaborati o trasferiti. La *Memory Protection Unit* fornisce una protezione agli accessi alla memoria.

Opzionale è l'unità di correzione da errori ECC che può rivelarne due, mentre gli errori di un singolo bit sono automaticamente corretti dal processore. La massima frequenza di clock è 600 MHz circa. L'area totale di 0,5 mm^2 e le prestazioni raggiungono circa 1000 DMIPS.

- ARM CORTEX-R7

  Estende le funzionalità del Cortex-R4.

  Il processore è dotato di due core ad alte prestazioni, per un largo campo di applicazioni profondamente integrate. Rispetto agli altri processori della stessa famiglia raggiunge un più alto livello di performance attraverso l'introduzione di una nuova tecnologia che include l'esecuzione di istruzioni fuori ordine e la rinominazione dinamica dei registri. Tutto questo è combinato con un miglioramento nella predizione dei salti, capacità di esecuzione superscalare ad 11 pipeline e supporto hardware più veloce per operazioni di divisione, DSP (*Digital Signal Processor*) e virgola mobile.



Struttura del processore Cortex-R7. [9]

Attualmente il processo ha un passo di 28 nm, quindi ad alta densità, inoltre utilizza celle di librerie a prestazioni standard. Le cache sono entrambe di 32 kbyte.

La frequenza massima di clock è circa 1 GHz e l'area totale di un singolo chip non supera i 0,7 mm^2. Le prestazioni raggiungono i 2500 DMIPS.

4.1.3 Considerazioni

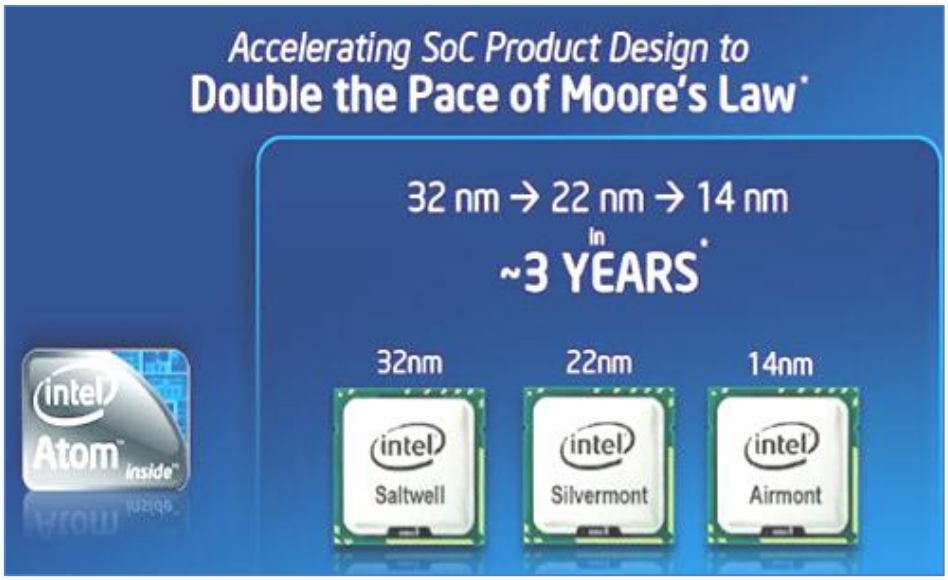
Osservando le caratteristiche dei processori la famiglia Cortex-A è quella con le più elevate prestazioni.

Inoltre i dispositivi di questa famiglia si possono presentare oltre alla versione single-core anche in modalità dual-core e quad-core. I Cortex-A sono ampiamente utilizzati nei dispositivi di ultima generazione ad elevate prestazioni: smatphone, netbook, tablet, sino ai server di ultima generazione. I Cortex-R sono invece ampiamente richiesti nei semiconduttori, circuiti integrati e microcontrollori di diverse case produttrici. Questi semiconduttori e microcontrollori vengono poi implementati in svariate applicazioni: settore automobilistico, settore industriale e dell'automazione, settori dell'elettronica di consumo.

Riguardo i consumi energetici il Cortex-A5 è il processore con il minor consumo energetico in configurazione single-core. Se consideriamo invece i Cortex-A5 e Cortex-A9 in configurazione dual o quad core i consumi energetici sono nettamente superiori ai processori Cortex-R.

4.2. Intel Atom

I processori Atom sono sviluppati da Intel con architettura x86, e sono destinati espressamente per il settore dei dispositivi portatili e notebook/desktop di fascia molto economica.
L'architettura x86 indica l'architettura introdotta da Intel, attualmente la più diffusa nel mercato dei PC desktop, portatili, e server economici. L'unico concorrente di Intel di un certo livello ad utilizzare questa configurazione è AMD.
Processori che implementano l'architettura x86 usano istruzioni di tipo CISC che permettono di risparmiare memoria, ma complicano notevolmente il progetti di nuovi processori.
Il primo SoC di di questa famiglia ad essere svelato fu Atom Z670 con un processo che da 45 nm volse poi ad uno con passo di 32 nm. I consumi di potenza sono calati insieme alle dimensioni, ma non abbastanza: Intel è ancora lontana da competere con ARM Holdings nel mercato di tablet e smartphone.
L'annuncio fatto da Intel per il futuro è Silvermont (sempre della famiglia Atom), un processore costruito con processo a 22 nm che dovrebbe dare ad Intel un vantaggio considerevole, consumando il 30 per cento di energia in meno rispetto ai suoi predecessori.
Intel non ha rivelato circa la progettazione di Silvermont, ma si vocifera in un miglioramento delle prestazioni negli acceleratori: dispositivi già in uso da Intel in grado di spezzare i programmi in più pezzi che possono essere eseguiti in parallelo. Questo consentirebbe di aumentare prestazioni, e risparmiare energia.
Il costo di molti dispositivi portatili è inferiore rispetto a quello di molti notebook. Giganti nell'industria del portatile inoltre hanno anni di esperienza nell'abbassare i costi con quote di mercato considerevoli per compensare i margini ristretti. Tutto ciò gioca a svantaggio di Intel.
Dal canto suo Intel ha qualcosa che manca ai concorrenti: l'integrazione verticale. Intel progetta e produce i suoi stessi chip, mentre altre aziende devono pagare qualcuno per farlo.



Panoramica evolutiva della tecnologia dei processori Intel Atom.

I concorrenti di Intel potrebbero affrontare un'altra minaccia ancora più grande nel 2014, quando Intel prevede di spostare tutti i suoi chip per PC, server, tablet e telefonini ad un processo produttivo a 14 nm. I costi saranno elevati, ma i guadagni forse ancora più ampi se l'azienda produce a due generazioni più avanti rispetto agli altri.

L'evoluzione prevista è descritta dall'immagine.

Partner commerciali di Intel per application processors per ora sono Motorola e Lenovo, multinazionale fondata in Cina.

Intel potrebbe diventare un avversario temibile nella corsa alle vendite dei processori per applicazioni negli anni a venire.

5. Exynos 4

I palmari di ultima generazione eseguono funzioni più svariate: di telefonia, di riproduzione musica e video, di fotocamera, di navigazione, di periferica di gioco ecc. Essi richiedono dunque un elevato *throughput* (la quantità di istruzioni eseguite in una data quantità di tempo) dei dati e risposte veloci nelle applicazioni multimediali tra cui giochi e navigazione web; un display ad alta risoluzione con un interfaccia intuitiva per una migliore interazione con l'utente; bassi consumi per estendere il più possibile la vita della batteria.

In questo capitolo verrà introdotto Exynos 4, un SoC per applicazioni embedded di casa Samsung costruito con un processo a 32 nm, quad-core.

Questo nucleo viene utilizzato come processore nello smartphone di penultima generazione (lanciato nel mercato a maggio 2012) Samsung Galaxy SIII.

Non ha bisogno di presentazioni.

La tecnologia di produzione dei semiconduttori ha positivamente risentito del continuo *scaling* di maschere ottiche ed in generale delle dimensioni geometriche dei dispositivi. Ad ogni migrazione di processo, le prestazioni, la potenza, il livello di integrazione seguiva la legge di Moore. Ma dal processo CMOS con passo di 90 nm il tradizionale scaling geometrico ha aperto sfide tecnologiche per sopperire la degradazione della mobilità, le resistenze parassite, le perdite di gate. Per mantenere i livelli di performance e i consumi desiderati dello scaling tradizionale, è stato necessario introdurre nuovi materiali nel processo dei semiconduttori come rame o isolanti high-k (materiali con elevata costante dielettrica).

Questo AP da 32 nm utilizza un processo high-k *metal gate* (HKMG) per minimizzare le perdite e massimizzare lo spettro di applicazioni. Dato processo consuma il 20 per cento in meno della precedente generazione di processori Exynos da 45 nm [9].

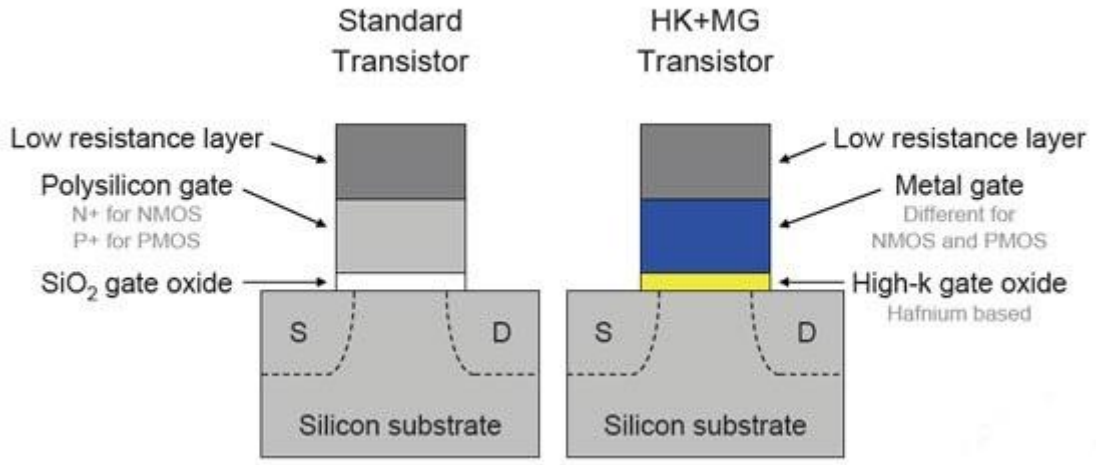
5.1. HKMG

Bassi consumi sono un obbiettivo primario da raggiungere nell'odierna progettazione di dispositivi portatili. I vincoli in potenza sono il problema più citato dai progettisti di semiconduttori e questo trend non può che continuare nella battaglia per energia pulita, e la continua domanda per piccoli fattori di forma con batterie di lunga durata.

L'innovazione del high-k metal gate come rimpiazzo del tradizionale Poly/SiON gate ha rotto le barriere dello scaling tecnologico fornendo significativi vantaggi in termini di potenza e prestazioni, garantendo il rimpicciolimento geometrico desiderato.

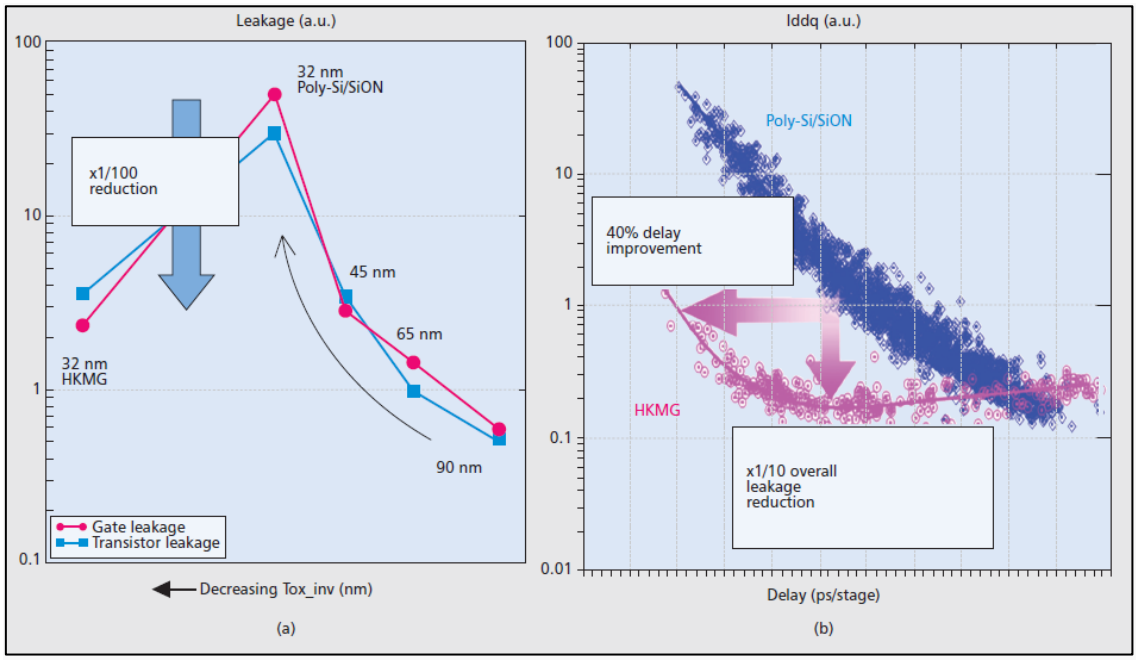
Il processo tradizionale Poly/SiON va incontro a limitazioni: come la tecnologia avanza, soffre di un esponenziale incremento di perdite e abbassamento delle prestazioni.

HKMG supera la sfida rimpiazzando l'ossido di silicio (usato come isolante di gate) con materiale high-k che riduce le perdite permettendo lo scaling dello spessore dell'ossido di gate. Un transistor cosi costruito si può osservare nell'immagine sottostante.



Costituzione di un transistor HKMG. Il gate metallico serve a prevenire una eventuale capacita parassita tra il materiale high-k e le piste di interconnessione, avente come dielettrico il polisilicio del gate.

Il processo a 32 nm HKMG può raggiungere una riduzione di un fattore 100 nelle perdite di gate e di un fattore 10 nelle perdite totali; inoltre le prestazioni aumentano del 40 % allo stesso livello di correnti di perdita. Come si vede dalla figura più a destra il ritardo è più alto per un gate in polisilicio standard.



Caratteristiche del high-k metal gate in termini di perdite; (a) corrente di perdita; (b) ritardo nel gate.

5.2. Il processore e i maggiori blocchi funzionali

Questo SoC integra un'architettura ARMv7 nota: Cortex-A9.

La CPU è dunque costituita da quattro core e similarmente anche la GPU (*graphics processing unit*) è composta da quattro *pixel-processor*. Sono presenti numerosi acceleratori multimediali ed altri blocchi per permettere la connettività e l'interfaccia con l'utente; così come il controllore DRAM (*dynamic* RAM) a due porte. Nell'immagine il diagramma a blocchi del processore.

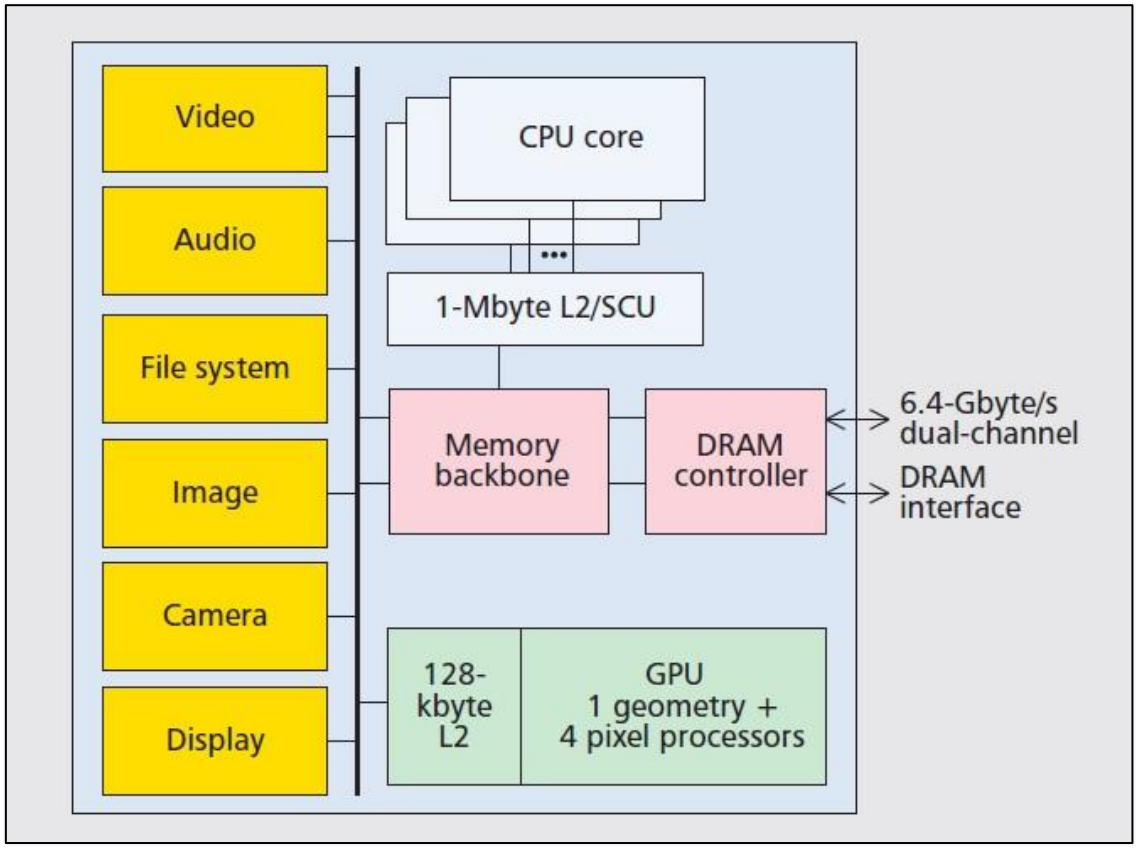


Diagramma a blocchi dell'architettura. "File System" indica il blocco di I/O che consente l'utilizzo di porte USB, eMMC (sono le schede di memoria antecedenti le SD) o SD/µSD.

Ogni CPU implementa una floating point unit ed una 64 bit ARM NEON single processor multiple data, come previsto per Cortex-A9.

Ogni core condivide una cache L2 da 1 Mbyte e ha una interfaccia sincrona con *snoop control unit* (SCU) che mantiene i dati delle cache coerenti tra i vari processori, inizializza l'accesso alla memoria, risolve le eventuali collisioni tra richieste di più core e organizza ACP (*Accelerator Coherency Port*).

La GPU processa 57 Mpolygons/s e supporta OpenGL ES 1.1/2.0 usando un processore di pixel quad-core, un *geometry processor*, una cache dedicata da 128 kbyte L2.

Il controllore DRAM e costituito da porte a 6,4 Gbytes/s con interfacce per memorie LPDDR2 (*low power* DDR2), DDR2 e DDR3.

Una delle sfide nel progetto della CPU è stato minimizzare l'accesso alla memoria; a questo proposito si è mantenuta una interfaccia sincrona a singolo ciclo tra i core e la SCU, sfruttando il progetto a singolo chip multicore.

Una interfaccia sincrona è difficile da raggiungere in progetti multichip e anche in soluzioni di un chip ma questa aiuta direttamente a minimizzare l'accesso dalle CPU alla cache L2.

La sincronia limita la velocità del core attivo, costringendo tutti a funzionare con la stessa frequenza di clock, ma aiuta ad evitare l'eccessivo tempo di latenza di core a bassa velocità.

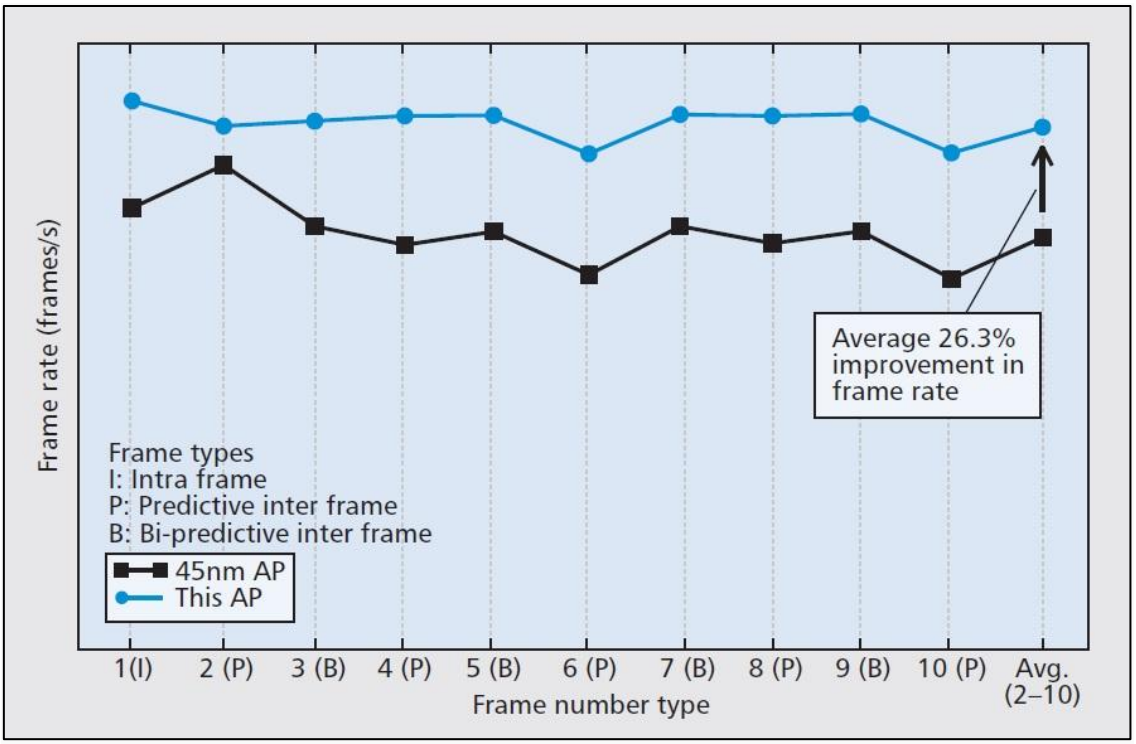
Per una configurazione quad-core un semplice riuso della architettura di un core non può raggiungere un interfaccia sincrona a singolo ciclo.

Il flusso di dati è organizzato gerarchicamente nell'architettura del bus puntando all'efficienza energetica, evitando connessioni *point-to-point* che consumino troppa potenza.

Per evitare l'intrinsecamente bassa larghezza di banda della struttura gerarchica del bus, il controllore DRAM supporta lo scambio fino a 32-byte di dati tra i due canali di memoria ed è sostenuta da un protocollo di interscambio grande 128 bit a 6,4 Gbytes/s.

Il sottosistema di gestione della memoria video è inoltre ottimizzato con una unità di controllo di memoria che supporta un indirizzo a due livelli che cattura con un largo TBL (*Translation Lookaside Buffer*). Copre così il grande spazio di indirizzo della memoria richiesto per grandi insiemi di dati multimediali. La combinazione di queste funzionalità forniscono un prestazioni di cattura video per 1080p con un anteprima LCD e uscita TV superiore in media del 26,3 % se comparato con la precedente tecnologia (45 nm, Samsung Galaxy SII uscito nel mercato a febbraio 2011).

Nella figura che segue è possibile osservare proprio le prestazioni nella cattura video in frame al secondo, in base anche al tipo di cattura che si sta eseguendo.



Miglioramenti delle prestazioni di cattura video comparate con un processore per applicazioni Samsung costruito con una tecnologia di 45 nm (Samsung Galaxy SII). Cattura con anteprima LCD.

L'architettura delle interconnessioni è progettata per minimizzare il ritardo di accesso alla memoria della CPU, infatti diversamente dagli altri blocchi funzionali la CPU è connessa direttamente a questa, garantendo un'alta priorità e bassi tempi di attesa.

Altre funzionalità sono distribuite in una gerarchia di bus in base alle esigenze ed ai ritardi. Evitare l'interfaccia asincrona a ciclo multiplo consente di diminuire i ritardi di lettura e scrittura della CPU dell'11 e 18 per cento rispettivamente.

5.3. Gestione dell'alimentazione

I recenti dispositivi portatili tendono ad avere sempre maggiori prestazioni e funzionalità, questo costringe a dover integrare sempre più transistor nel processore ed a imporgli una frequenza maggiore di funzionamento.

Ciò porta a consumare più potenza dinamica durante le operazioni in attività, e perdere grandi quantità di corrente durante gli stati inattivi o *sleep state*. Questo minaccia la portabilità dei dispositivi riducendo la vita della batteria.

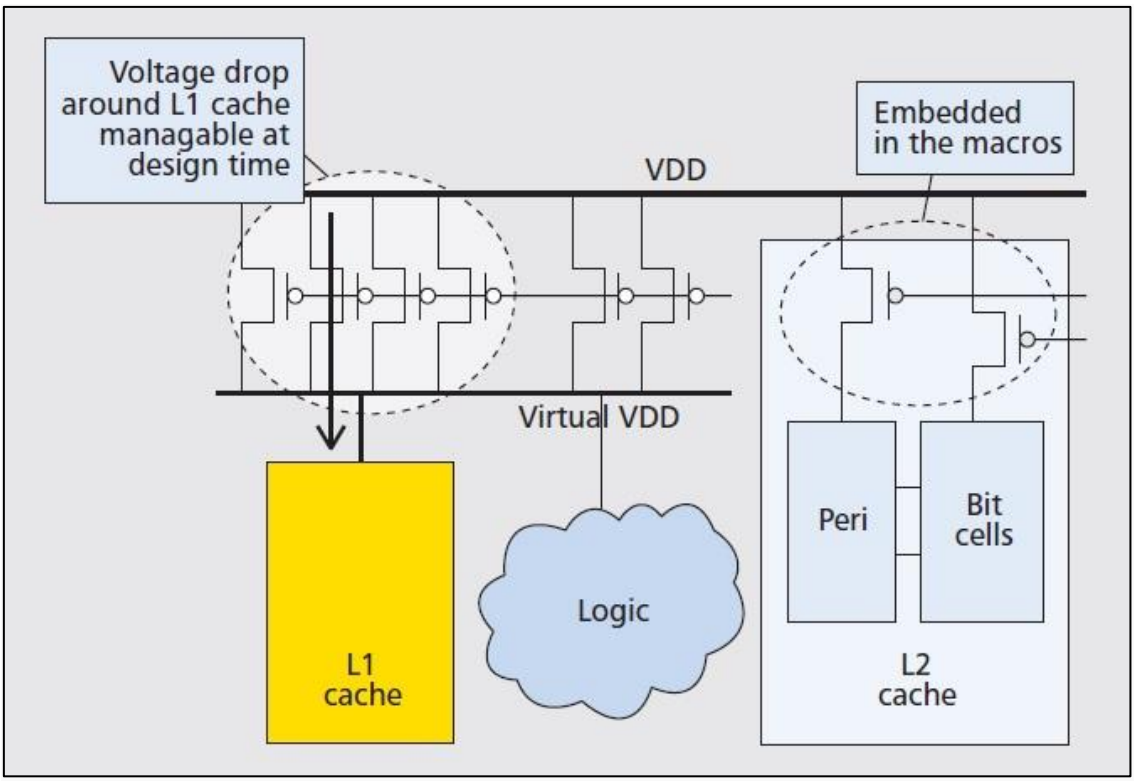
In aggiunta alla tecnologia di processo per bassi consumi, la gestione di potenza a livello di sistema diviene quindi molto importante per fornire la giusta quantità di potenza dinamica per le prestazioni richieste e per minimizzare le perdite di corrente negli stati inattivi.

Per organizzare il consumo di energia ed evitare sprechi questo AP ha un'unità di gestione della potenza, il PMU (*power management unit*). Il PMU controlla accensione e spegnimento del power gating interno ed opera in tensione/frequenza per gestire rispettivamente perdite e potenza dinamica.

Ci sono un numero di diversi piani di funzionamento tra cui quattro per la CPU, GPU, DRAM, e gli altri blocchi di funzioni multimediali. La tensione di funzionamento e la frequenza sono dinamicamente ed indipendentemente controllate dalla PMU, il clock è basato sul carico di lavoro e sullo stato termico analizzato dal software della CPU. Questo controlla il consumo di potenza dinamica durante lo stato di attività. Ogni piano di funzionamento ha più indipendenti domini di potenza per il power gating e consente un preciso controllo delle perdite nei consumi mentre sono inattivi.

Tutte le CPU dei core e la loro cache condivisa L2 appartiene alla stesso piano di potenza, che è scalato tra 200 MHz e 1,6 GHz. Ogni core e la cache possono essere gestiti indipendentemente in base alla produttività del processo e all'ammontare di dati richiesti dall'applicazione. L'idea di base è di far andare la PMU con un primo semplice schema di clock: questo controlla la frequenza dei core contemporaneamente, e quindi di tutti i core uno per uno quando questi sono ad un livello di frequenza più basso del necessario per le performance richieste. Quando arriva una richiesta improvvisa di prestazioni la CPU deve passare dallo stato in cui tutti i core sono spenti, allo stato in cui tutti lavorano alla massima frequenza, e questo per minimizzare i tempi di risposta. Inoltre la PMU può essere programmata con schemi più complicati in base all'uso da farsi.

Dalla parte dell'implementazione, le cache L1 e L2 hanno differenti strutture per il power gating: la cache L2 ha un sistema integrato nella memoria per massimizzare la densità; L1 invece è implementata con le celle per il power gating all'esterno di essa, questo permette al programmatore di gestire accuratamente la tensione in L1 e minimizzare l'impatto delle sue prestazioni. In dettaglio nell'immagine che segue.



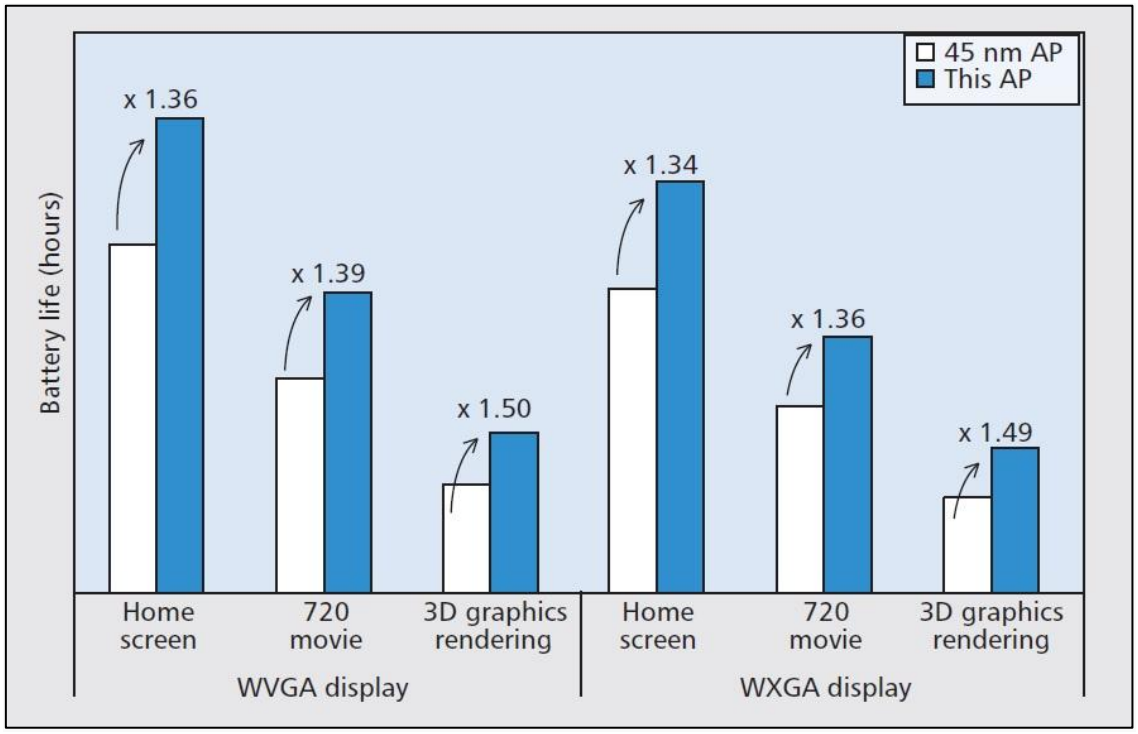
Struttura per il power gating delle cache L1 e L2.

Due piani specifici sono dedicati per il controllo della potenza nella GPU e nei blocchi multimediali, rispettivamente. La GPU è uno dei blocchi più grandi nei processori per applicazioni con un enorme variazione in carico di lavoro. Non viene impiegata semplicemente una interfaccia utente semplice o di navigazione web 2D, ma una complicata grafica 3D con un alto traffico di dati come sono i giochi.

Quindi la GPU è alimentata dal suo proprio piano di potenza per permettere ottimizzazioni indipendenti in tensione/velocità e ha un sistema di gestione locale di potenza a livello di core per il clock/power gating. Questo schema di gestione della potenza è incorporato con l'interfaccia di programmazione grafica (API) e ottimizzato a livello di software. Gli acceleratori multimediali includono i sottosistemi audio e video, e usano un altro piano di potenza condiviso tramite bus interni.

Tra tutte le funzionalità multimediali, la riproduzione audio è una delle più frequenti e più utilizzate nei recenti dispositivi portatili. Quindi per permettere bassissimi consumi di potenza e tempi in riproduzione extra lunghi, il sottosistema audio include un processore dedicato riconfigurabile e canali audio dedicati. Questi permettono controlli indipendenti del power gating e la PMU accende un minimo insieme di funzioni e interconnessioni ad un livello predefinito di tensione di alimentazione.

I test con i primi campioni hanno verificato che queste configurazioni combinate con il processo tecnologico HKMG migliorano significativamente il consumo di potenza.



Vita della batteria del dispositivo in esame comparata con il processo Samsung a 45 nm del suo predecessore.

Come si può osservare nell'immagine, il nuovo processore in attesa sulla schermata iniziale dura il 34-36 per cento più a lungo con la stessa capacità di batteria (dipende comunque dalle dimensioni del display) rispetto al processore con la precedente tecnologia. Per applicazioni con un carico di dati elevato, come film a 720p e riproduzione grafica 3D, si nota un aumento in percentuale della durata della batteria di 36-39 e 49-50, rispettivamente.

Un altro aspetto da tener però presente è che le ultime tecnologie rendono più difficile la gestione delle perdite di potenza. Questo è dovuto alla crescente variazione di processo. Pertanto ridurre la variazione delle finestre è la chiave per la corretta organizzazione e il controllo delle perdite di corrente attorno ad un livello desiderato.

Per arginare questo problema, in questo processore sono distribuiti un numero di controllori per determinare le perdite e le prestazioni, e per catturare le caratteristiche del die (piastrina di semiconduttore nel quale è realizzato il circuito integrato).
I dati dei controllori sono analizzati per identificare le "spigolature" di processo, il valore delle tensioni di soglia e le altre variazioni nel chip; dopo di ché vengono utilizzati per il controllo della polarizzazione e della tensione per i piani di alimentazione. Questo consente di ridurre la finestra dei campioni in silicio e minimizzare l'impatto di perdite/prestazioni delle variazioni.
Alcuni chip usati per le analisi e per il progetto del dispositivo finale sono più interessanti di altri. Questi die hanno caratteristiche che spaziano da un estremo all'altro nella tecnologia utilizzata; in questo caso i chip in silicio utili all'analisi sono gli estremi di processo (*process corners*):

- SS (NMOS lenti, PMOS lenti) che minimizza la perdita totale in corrente, e consente di organizzare in modo selettivo i blocchi più lenti.

- FF (NMOS veloci, PMOS veloci) per la gestione dei blocchi con dispersione critica.

L'utilità di questi risiede nel consentire ai progettisti di rilevare effetti indesiderati (che possono avvenire con probabilità non nulla) ed evitare quindi perdite o lentezza eccessiva in alcuni blocchi del sistema, prima che il design sia strutturato e mandato in produzione definitiva.

5.4. Dissipazione del calore/Thermal Management Unit

Come descritto fin'ora i dispositivi portatili consumano più energia di una volta ma non possiedono sistemi di raffreddamento delicati e costosi presenti nelle macchine di calcolo ad alte prestazioni.
D'altra parte un sistema real-time di questo tipo deve garantire una risposta in tempo finito senza ritardare, attivando per esempio protocolli per il raffreddamento che inibiscano l'utilizzo dello smartphone all'utente per tempi troppo lunghi.
Piccoli fattori di forma e bassa capacità di raffreddamento sono uno svantaggio per la dissipazione di calore. In aggiunta, i recenti fornitori di smartphone hanno degli standard da seguire per mantenere le superfici dei dispositivi al di sotto di una certa temperatura in modo da garantire tutto il confort possibile dell'utilizzatore. Principalmente per questo motivo e per evitare malfunzionamenti del dispositivo dovuti al calore, è installata una unità di gestione termica (TMU).

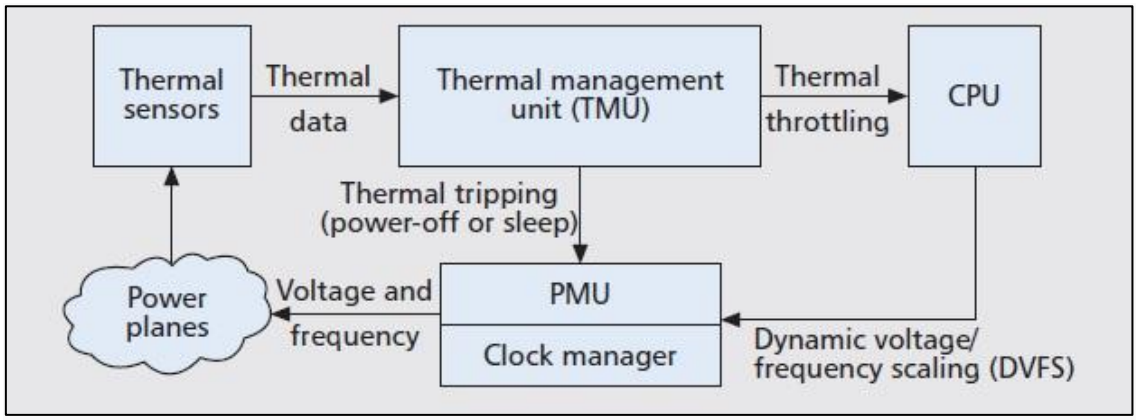
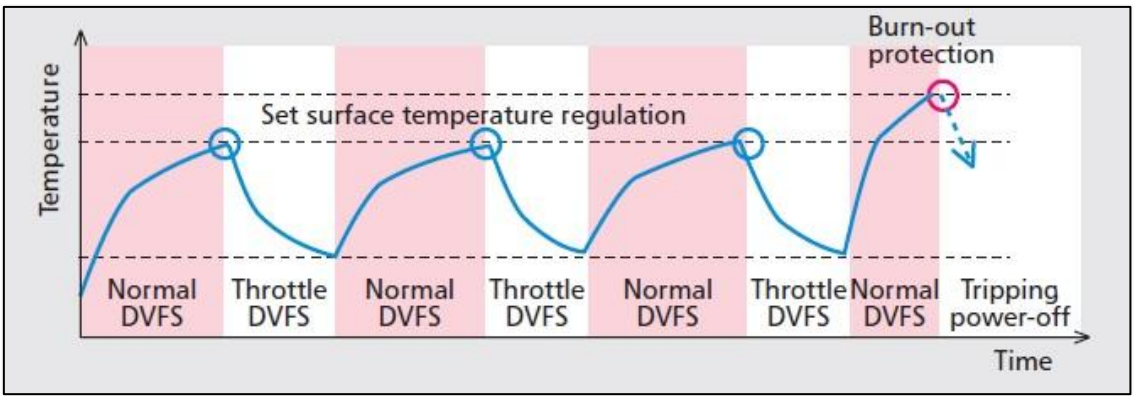


Diagramma a blocchi e principio di funzionamento attorno all'unità di gestione della temperatura TMU.

Dallo schema a blocchi in figura è chiaro il principio di funzionamento della TMU: un sensore cattura accuratamente la temperatura come campione, altri sensori distribuiti nel die misurano la temperatura relativa determinando i "punti caldi". La TMU agisce attivando limitatori che tengono le temperature sotto la soglia prestabilita: quando un sensore invia l'informazione di un punto troppo caldo, tensione e frequenza sono dinamicamente scalate dal software (firmware) della CPU.

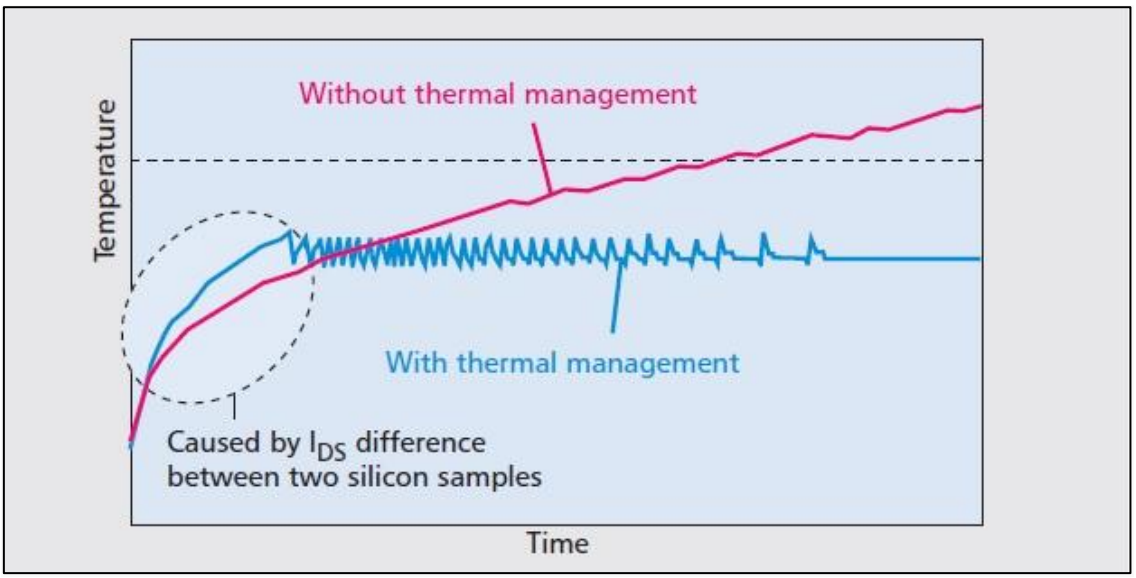
Quando il dispositivo degrada termicamente e la temperatura esplode significa che il software di controllo è troppo lento. Dunque, per proteggere il processore da un guasto, la TMU immediatamente spegne l'alimentazione dei maggiori blocchi attraverso la PMU finché la temperatura non ritorna a livelli accettabili. Un esempio nella figura sottostante.



Esempio di gestione di un aumento della temperatura oltre la soglia.

Alcuni test sono stati effettuati per analizzare il comportamento della temperatura in chip campioni, con i maggiori blocchi funzionanti per il massimo delle prestazioni (decodifica video 1080p; giochi istantanei 2D nella CPU; prestazioni grafiche 3D nella GPU).

La TMU mantiene le temperature sotto una data soglia mentre senza TMU la temperatura aumenterebbe molto di più come si può osservare nell'immagine sottostante.



Dispositivi con e senza TMU. Da notare quanto, senza l'unità di controllo della temperatura, l'aumento di questa nel sistema sia progressivo e costante.

6. Conclusioni

I dispositivi portatili che siamo abituati a utilizzare tutti i giorni contengono al loro interno le tecnologie più avanzate, per garantire all'utilizzatore la possibilità di reperire una grande quantità di informazioni, ovunque esso si trovi. Infatti quasi ogni cellulare ormai garantisce l'accesso ad internet.

Non solo, grazie al sistema di ricezione GPS integrato nella maggior parte degli smartphone non è più necessario circolare con piantine e spesso nemmeno chiedere informazioni per raggiungere luoghi di cui non conosciamo la strada.

Queste due operazioni che ci viene spontaneo ormai fare in caso di bisogno, sono solo alcuni dei servizi che ci fornisce il nostro dispositivo mobile, oltre a fotocamera, lettore musicale, console, servizi di telefonia, ecc..

Fino a dieci anni fa ciò che sorprendeva era solamente la possibilità di telefonare ovunque ci trovassimo, con la diffusione dei primi cellulari. L'evoluzione è stata sorprendente, ed ha cambiato le nostre vite.

Il mio contributo per ora è stato scrivere questo documento in cui ho messo in luce più aspetti possibile delle caratteristiche che deve avere il nucleo di questi sistemi: il microprocessore.

Infatti è proprio questo elemento che coordina e rende possibile la convivenza di così tante funzioni così diverse.

Sono molto soddisfatto di aver appreso come siano costituiti e quali problemi comporta il progetto di un processore per applicazioni real-time a basso consumo come quello di un sistema portatile di ultima generazione.

I motivi sono principalmente due, il primo è che il mercato di questi dispositivi è in continuo aggiornamento e occupa una fetta molto consistente nell'ambito dei dispositivi elettronici digitali.

Il secondo è di sentirmi fortunato a conoscere cosa voglia dire produrre un oggetto di questo livello assieme anche agli anni di lavoro che sono stati fatti per raggiungere certe tecnologie;

e poter quindi apprezzare interamente il lavoro dell' uomo e del suo intelletto semplicemente mettendo la mano in tasca per prendere il mio smartphone.

7. Riferimenti bibliografici

1. Linda Sui, "Smartphone Revenues, ASPs & Price-Tier Forecast:2009 to 2017", *http://www.strategyanalytics.com/default.aspx?mod=reportabstractviewer&a0=8578*, Giugno 2013.

2. Rachel Courtland, "The Intel-ARM Core War", *http://spectrum.ieee.org/tech-talk/semiconductors/design/the-intel-arm-core-war*, Gennaio 2012.

3. Katherine Bourzac, "Intel Inside…Your Smartphone", *http://spectrum.ieee.org/semiconductors/processors/intel-insideyour-smartphone*, Dicembre 2012.

4. Henry Blodget and Eleanor Miller, "Welcome To The Future Of Media [IGNITION DECK]", *http://www.businessinsider.com/the-future-of-media-2011-12?op=1*, Dicembre 2011.

5. O'Connor, Patrick D. T., *Practical Reliability Engineering* (Fourth Ed.), John Wiley & Sons, New York, 2002

6. Jayanth Srinivasan, Sarita V. Adve, Pradip Bose, Jude A. Rivers; "The Case for Lifetime Reliability-Aware Microprocessors", *The 31st International Symposium on Computer Architecture* (ISCA-04), June 2004

7. ARM Holdings, "Cortex A-series", *http://www.arm.com/products/processors/cortex-a*; "Cortex R-series" *http://www.arm.com/products/processors/cortex-r*

8. IndiaTek, "ARM Cortex A5 Processor Overview", *http://indiatek.wordpress.com/2012/04/14/arm-cortex-a5-processor-overview/*, Aprile 2012

9. Samsung Corporation, "Samsung Exynos 4 quad", *http://www.samsung.com/global/business/semiconductor/minisite/Exynos/products4quad.html*

10. Se-Hyun Yang, Jungyul Pyo, Youngmin Shin, and Jae Cheol Son, Samsung Electronics; "A 1.6 GHz Quad-Core Application Processor Manufactured in 32 nm High-k Metal Gate Process for Smart Mobile Devices", IEEE Communications Magazine, Aprile 2013

11. Kyu-Myung Choi, "32 nm High K Metal Gate (HKMG) Designs for Low Power Applications", International SoC Design Conference, 2008

12. Jan M. Rabaey, Anantha Chandrakasan, Bora Nikolic'; *Circuiti Integrati Digitali*, Pearson Education Italia, Milano, Settembre 2005

13. Sergio Congiu, *Architettura degli Elaboratori*(Quinta edizione), Pàtron Editore, Bologna, Marzo 2007