Stima di potenza nella progettazione di circuiti digitali a basso consumo

Alessandro Bogliolo*, Luca Benini*, Giovanni De Micheli** e Bruno Riccò*

* DEIS - Università di Bologna - Bologna - Italia
** CSL - Stanford University - Stanford - California

Sommario

Il progetto di circuiti integrati digitali a basso consumo richiede la disponibilità di stime di potenza accurate ad ogni livello di astrazione. Ma l'accuratezza delle stime decresce molto rapidamente all'aumentare del livello di astrazione. In questo lavoro descriviamo un simulatore di potenza gerarchico (denominato PPP) basato su modelli simbolici che superano i limiti di accuratezza propri dei livelli di astrazione ai quali operano grazie ad informazioni collezionate ai livelli di astrazione inferiori.

1. Introduzione

Il basso consumo di potenza è un requisito fondamentale per gli attuali circuiti integrati. Particolarmente critiche sono le esigenze dettate da applicazioni portatili (*laptop* e telefonia personale), in cui il consumo di potenza determina non solo l'affidabilità, ma anche la disponibilità (tempo di utilizzo) e la praticità d'uso (dimensioni e peso delle batterie) dell'apparato.

Il progetto di sistemi digitali a basso consumo di potenza avviene a diversi livelli di astrazione [1]. Ad ogni livello molti gradi di libertà devono effere fissati e svariati accorgimenti possono essere adottati per rispettare le specifiche di progetto. Stime di potenza rapide ed accurate sono necessarie per guidare e validare ogni scelta progettuale. Ma efficienza e accuratezza sono esigenze intrinsecamente contrastanti, e gli strumenti in grado di trattare circuiti di dimensioni realistiche forniscono stime fortemente approssimate. Del resto, il compromesso opposto che si realizza ai livelli di astrazione più bassi consente di analizzare solo singoli sottocircuiti di piccole dimensioni ed è praticamente inutilizzabile ai fini del progetto e della verifica di interi circuiti.

In questo contesto, il lavoro illustra un simulatore di potenza gerarchico (denominato PPP) che è stato concepito come supporto al progetto e all'analisi di circuiti basati su librerie di celle CMOS di cui sia nota l'implementazione (*gate* a livello logico, o primitive arbitrariamente complesse a livello *behavioral*). La strategia di simulazione si basa sull'uso di modelli avanzati degli elementi di libreria, che ad ogni livello di astrazione consentono di tener conto di effetti normalmente visibili solo a livelli di astrazione inferiori. Questo permette di realizzare nuovi compromessi tra efficienza ed accuratezza.

A livello logico vengono modellati non solo i fenomeni di carica e scarica delle capacità all'uscita di ogni *gate*, ma anche correnti di corto circuito, ridistribuzioni di carica interna e transizioni spurie (*glitch*) dovute alla commutazione non perfettamente simultanea di più segnali. A livello *behavioral* viene modellata la dipendenza del consumo di potenza dai vettori applicati all'ingresso del circuito.

Il modello di potenza di ogni elemento di libreria è caratterizzato una volta per tutte sulla base di simulazioni effettuate al livello di astrazione immediatamente inferiore: *HSPICE* è utilizzato per caratterizzare il model-

lo *gate*, mentre lo stesso simulatore logico di PPP è utilizzato per caratterizzare i modelli *behavioral*. La costruzione e la caratterizzazione dei modelli avvengono in modo completamente automatico.

2. Livello *gate*

La rappresentazione logica di un circuito si basa su una duplice astrazione: la prima è legata alla granularità della rappresentazione, che impedisce di apprezzare la struttura interna dei *gates*, la seconda è legata al concetto di segnale digitale, che impedisce di apprezzare la natura intrinsecamente analogica e tempo-continua di correnti e tensioni. Per poter effettuare stime di potenza occorre aggiungere alla descrizione logica informazioni di natura elettrica. La rappresentazione di informazioni di supporto che non sono proprie del livello di astrazione al quale si opera prende il nome di *backannotation*.

Le sole informazioni di natura elettrica generalmente utilizzate per la valutazione del consumo di potenza a livello logico sono la tensione di alimentazione (V_{dd}) e la capacità di carico di ogni *gate* (C_L). Queste consentono di stimare la potenza necessaria a caricare le capacità di carico dei *gates* che commutano:

$$P = \sum_g \frac{a_g}{2} \frac{V_{dd}^2 C_{Lg}}{T_{ck}} \qquad (1)$$

dove a_g è la probabilità media di transizione (o attività) del *gate* g e la sommatoria si intende estesa a tutti i componenti del circuito. La (1) ha il vantaggio fondamentale della semplicità, che ne permette tra l'altro la valutazione statica (senza bisogno di effettuare simulazioni) qualora siano note (o possano essere stimate) le probabilità di transizione dei segnali [2]. In tal modo, però, l'effettivo consumo di potenza può essere pesantemente sottostimato in quanto vengono completamente trascurati importanti fenomeni dissipativi quali correnti di corto circuito, carica e scarica di capacità interne e ridistribuzioni di carica. Inoltre, l'uso di modelli di ritardo

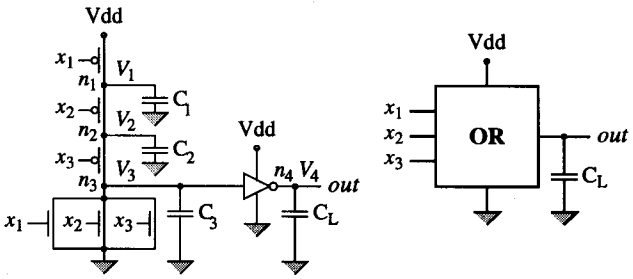

Figura 1: OR a tre ingressi in tecnologia CMOS. Gli accoppiamenti capacitivi sono rappresentati da quattro capacità costanti verso massa: $C_1 = C_2 = 11fF$, $C_3 = 157fF$ and $C_L = 136fF$.

semplificati (nullo, unitario o costante) per la propagazione dei segnali impedisce di apprezzare eventuali transizioni spurie (*glitches*) che possono aumentare di oltre il 20% il consumo di potenza totale [3].

Esempio 1. In Fig. 1 è rappresentata la realizzazione CMOS di un OR a tre ingressi. Dal punto di vista logico, la variazione della configurazione d'ingresso da $\mathbf{x} = 100$ a $\mathbf{x} = 010$ non produce alcun effetto (i caratteri in neretto sono utilizzati per rappresentare vettori: $\mathbf{x} = [x_1, x_2, x_3]$). Tuttavia, un disallineamento di 0.4ns tra i fronti di discesa e di salita di x_1 e x_2 dà luogo a un consumo di potenza non trascurabile dovuto a due fenomeni: 1) la doppia transizione del nodo d'uscita, che implica carica e scarica di C_L e C_3, e 2) la corrente di corto circuito attraverso entrambi gli stadi CMOS. Il corrispondente consumo di potenza stimato da HSPICE in un periodo di 20ns è di 0.08mW. □

In linea di principio, i limiti di rappresentazione del livello logico possono essere superati qualora il circuito sia mappato su una libreria di celle per le quali siano disponibili informazioni dettagliate (collezionate a livello elettrico) sul consumo di potenza [4]. La precaratterizzazione degli elementi della libreria è effettuata una volta per tutte e consiste in reiterate simulazioni a livello elettrico ripetute per ogni cella a fronte di ogni possibile transizione d'ingresso e per diverse condizioni di *fanin* e *fanout*. Le informazioni sul consumo di potenza fornite dalla simulazione elettrica vengono raccolte in tabelle (*look-up-tables*) associate alle celle. Durante la simulazione

logica, ogni volta che si verifica un evento in ingresso ad un *gate* (istanza di una cella di libreria) il corrispondente consumo di potenza è letto sulla tabella ad esso associata.

In teoria, se il valore del consumo di potenza di una cella potesse essere tabulato per ogni possibile transizione d'ingresso e per ogni valore dei parametri da cui esso dipende, la simulazione a livello logico raggiungerebbe l'accuratezza della simulazione elettrica usata per la caratterizzazione. Ma la precaratterizzazione non può essere esaustiva poichè la dimensione delle tabelle e il numero di simulazioni elettriche cresce esponenzialmente con il numero di parametri. Per ridurre la dimensione dei modelli e il tempo di caratterizzazione, nella pratica le simulazioni elettriche vengono effettuate solo per singole transizioni d'ingresso e per valori tipici dei parametri di *input/output*, e viene trascurata la dipendenza dallo stato di carica interno. L'approssimazione che ne deriva limita l'accuratezza delle stime di potenza.

Recentemente sono stati proposti approcci avanzati che sfruttano la conoscenza dei fenomeni dissipativi per semplificare la caratterizzazione e il modello del consumo di potenza di ogni cella. In [5] si osserva che il consumo di potenza manifesta due andamenti completamente diversi in funzione del rapporto tra la velocità di transizione dei segnali d'ingresso e d'uscita. Gli autori propongono quindi di usare due modelli diversi per approssimare il consumo di potenza nei due casi. L'accuratezza delle stime che ne derivano è però limitata dalla semplicità dei modelli adottati e dalla mancanza di informazioni dettagliate sui ritardi e sullo stato di carica interno. Lo stato di carica è invece alla base del modello proposto in [6], che rappresenta i *gates* CMOS come macchine a stati finiti, in cui il consumo di potenza è associato alle transizioni tra gli stati. Ma gli autori non propongono alcun modello analitico per esprimere la dipendenza del consumo di potenza dalla durata delle transizioni d'ingresso e dal valore della capacità di carico, e si avvalgono di *look-up-tables* associate ad ogni possibile transizione, ricadendo così nel compromesso tra accuratezza e complessità del modello. Infine, è importante osservare che nessuno degli approcci proposti in letteratura propone tecniche per modellare transizioni disallineate e *glitch*. Per superare i limiti appena discussi, il simulatore a livello *gate* di PPP si basa su un modello completamente simbolico.

2.1. Modello simbolico

In generale, la corrente $I(t)$ assorbita da un *gate* CMOS a seguito di una transizione d'ingresso può sempre essere vista come la somma di due contributi:

- $I_c(t)$ (*charging current*), corrente di caricamento che incrementa la quantità di carica immagazzinata sui nodi del circuito,
- $I_w(t)$ (*wasted current*), corrente che rifluisce direttamente verso massa senza modificare lo stato di carica del circuito.

Lo stesso partizionamento può essere applicato all'energia assorbita dal *gate* durante tutto il transitorio:

$$E_c = \int_{t^i}^{t^f} V_{dd} I_c(t) dt \quad E_w = \int_{t^i}^{t^f} V_{dd} I_w(t) dt$$

Gli apici i ed f indicano l'inizio e la fine di un transitorio provocato dalla transizione d'ingresso dal vettore $\mathbf{x}^i$ al vettore $\mathbf{x}^f$.
Benchè non sia facile (anche nell'ambito di una simulazione elettrica) distinguere I_w da I_c, l'energia di caricamento E_c può essere facilmente valutata in base allo stato di carica della cella all'inizio e alla fine del transitorio:

$$E_c = V_{dd} \int_{t^i}^{t^f} I_c(t) dt = V_{dd} \, \Delta Q_c$$

dove ΔQ_c è la quantità totale di carica fornita dall'alimentazione alle capacità parassite (interne e di carico). Del resto, nota E_c è banale ottenere E_w per differenza: $E_w = E - E_c$.

Si noti che E_c *non* dipende dalla pendenza dei fronti dei segnali e la sua valutazione richiede solo la conoscenza dello stato di carica (o dei livelli di tensione) ai nodi della cella. Viceversa, E_w *non* dipende dallo stato della cella e può essere espresso in funzione dei soli parametri d'ingresso e d'uscita. Senza perdita di accuratezza, il problema di modellistica può quindi essere partizionato in due sottoproblemi più semplici il cui scopo è quello di modellare separatamente E_c ed E_w.
Benchè non sia possibile in questo contesto illustrare in dettaglio i modelli simbolici di E_c e E_w (si veda [7] a tal fine), ne riassumiamo le caratteristiche principali. Come premesso, il modello di E_c si basa sulla determinazione delle variazioni di carica ai nodi:

$$\Delta Q_c = \sum_{n_i \in \mathcal{S}} \Delta q_i \qquad (2)$$

dove $\mathcal{S}$ è l'insieme dei nodi collegati a V_{dd} a fronte del vettore d'ingresso $\mathbf{x}^f$. Poichè le condizioni di attivazione dei cammini conduttivi interni ai *gates* possono essere espresse da funzioni booleane degli ingressi, le variazioni di carica interna possono essere calcolate in modo simbolico durante la simulazione logica senza richiedere alcuna approssimazione. Il modello adottato per E_w è invece una funzione lineare della pendenza dei fronti d'ingresso e della capacità di carico. I coefficienti del modello sono determinati durante la caratterizzazione in modo da minimizzare l'errore quadratico medio compiuto dal modello rispetto alla simulazione elettrica.

2.2. Transizioni multiple disallineate

Il modello simbolico di $E = E_c + E_w$ fornisce stime accurate (*pattern-dependent*) del consumo di potenza di celle CMOS a fronte di transizioni d'ingresso tra due vettori qualsiasi $\mathbf{x}^i$ e $\mathbf{x}^f$. L'ipotesi implicita è che se $\mathbf{x}^i$ e $\mathbf{x}^f$ differiscono per più di un bit, tutti i segnali che commutano lo facciano contemporaneamente. Nella pratica, tuttavia, questa situazione è poco realistica, poichè la propagazione attraverso cammini di lunghezza divera provoca sfasamenti difficilmente prevedibili tra i segnali.

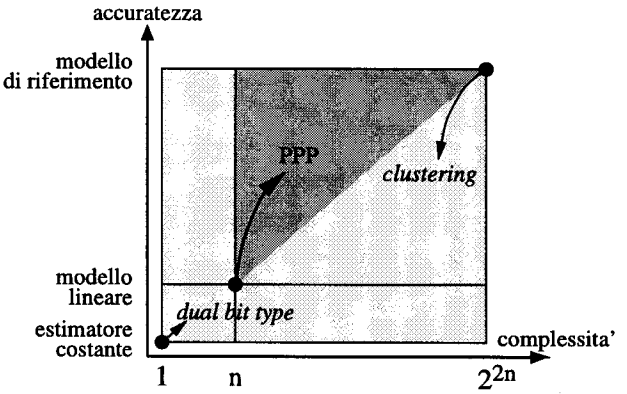

Figura 2: Rappresentazione qualitativa dei possibili compromessi tra accuratezza e complessità dei modelli di potenza a livello *behavioral*.

Per valutare il consumo di potenza di una cella a fronte di transizioni disallineate usiamo un metodo di interpolazione lineare tra due situazioni limite:

1. il caso in cui le transizioni siano perfettamente allineate,
2. il caso in cui le transizioni siano completamente disgiunte.

Poichè il modello descritto nel paragrafo precedente fornisce stime accurate di E in entrambe le situazioni, queste vengono assunte come estremi per l'interpolazione lineare che fornisce la stima di energia in tutti i casi intermedi.

3. Livello *behavioral*

La distanza tra la descrizione comportamentale di un circuito e i meccanismi fisici che ne determinano il consumo di potenza è tale da non permettere di apprezzarne il legame causale. Di conseguenza non è praticamente possibile partire dall'analisi dei fenomeni dissipativi per costruirne un modello comportamentale, e si utilizzano piuttosto modelli astratti il cui significato è esclusivamente statistico.
Sul piano complessità-accuratezza (rappresentato in Fig. 2) tutti i possibili modelli

sono racchiusi in un rettangolo ai cui estremi stanno il modello costante e il modello di riferimento. Il primo è il più semplice modello di potenza *pattern-independent*, e rappresenta la potenza media dissipata dal circuito durante gli esperimenti condotti in fase di caratterizzazione. Il secondo è il modello di riferimento *pattern-dependent*, che ad ogni possibile coppia di configurazioni d'ingresso ($\mathbf{x}^i$, $\mathbf{x}^f$) associa il corrispondente valore del consumo di potenza (stimato, generalmente, attraverso simulazioni a livello *gate*). La ricerca nel campo della modellistica del consumo di potenza a livello comportamentale si è mossa finora in due direzioni: sviluppando modelli *pattern-independent* in grado di aumentare l'accuratezza del modello costante senza aumentarne troppo la complessità; sviluppando modelli *pattern-dependent* che riducessero la complessità del modello di riferimento senza ridurne troppo l'accuratezza.

Alla prima categoria appartengono il modello *dual bit type* proposto da Landman e Rabaey nel '95 [8] e la maggior parte dei modelli di potenza a livello *behavioral*, mentre alla seconda categoria appartiene il metodo di *clustering* proposto da Mehta ed altri nel '96 [9]. Ma, in generale, l'accuratezza dei modelli *pattern-independent* risente pesantemente della distribuzione statistica dei vettori d'ingresso, mentre l'accuratezza del metodo di *clustering* è compromessa dall'ipotesi irrealistica su cui esso si basa: che a configurazioni d'ingresso simili corrispondano consumi di potenza simili.

I modelli di potenza utilizzati da PPP si collocano in una diversa regione del piano complessità-accuratezza, posta al di sopra del punto rappresentativo dei modelli lineari.

3.1. Modello lineare

L'andamento di $P(\mathbf{x}^i, \mathbf{x}^f)$ è tutt'altro che lineare e l'uso di approssimazioni lineari comporta pesanti errori, simili a quelli che si otterrebbero spingendo il criterio di *clustering* [9] verso semplificazioni estreme. In entrambi i casi, l'errore sta innanzitutto nel tentativo di approssimare la dipendenza di P da variabili di per sè poco significative. Benchè ogni coppia $(\mathbf{x}^i, \mathbf{x}^f)$ individui univocamente il consumo di energia del circuito, la dipendenza di P da ogni singola variabile è praticamente priva di significato fisico e fortemente correlata al valore delle altre variabili. La scelta di variabili più adatte ad esprimere in modo semplificato la *pattern-dependence* di P è suggerita da due considerazioni:

- in un circuito combinatorio CMOS qualche ingresso deve commutare perchè vi sia dissipazione di potenza,
- la presenza di transizioni in uscita indica la presenza di attività interna.

Alla luce di queste considerazioni esprimiamo il consumo di potenza come funzione lineare dell'attività dei segnali d'ingresso e d'uscita, piuttosto che dei loro valori. In simboli, il modello di potenza per un blocco con n ingressi ed m uscite è:

$$P = c_0 + c_1 a_1 + ... + c_n a_n + ... + c_{n+m} a_{n+m}$$

dove $\mathbf{c} = (c_0, c_1, ..., c_{n+m})$ è il vettore di coefficienti da determinare in fase di caratterizzazione, e $\mathbf{a} = (a_1, ..., a_{n+m})$ è il vettore di variabili indipendenti che assumono valore 1 in presenza di una transizione sulla corrispondente linea di ingresso o uscita.

La determinazione dei coefficienti avviene attraverso *fitting* ai minimi quadrati su un campione di dati raccolti durante la simulazione a livello *gate* del circuito da caratterizzare. I modelli lineari caratterizzati ai minimi quadrati producono stime di P con lo stesso valor medio della funzione di riferimento sul campione utilizzato per il *fitting*. Questo garantisce al modello lineare un'accuratezza comunque non inferiore a quella dell'approssimazione costante.

3.2. Modelli avanzati

Benchè i modelli lineari siano molto più accurati e flessibili del modello costante e siano

molto più efficienti del modello di riferimento, l'effettivo andamento del consumo di potenza è fortemente non lineare. Tentare di approssimarlo con una funzione lineare può comportare gravi errori la cui entità dipende dalla significatività del campione di dati utilizzato per la caratterizzazione. Tanto più questo è rappresentativo del reale contesto operativo del circuito, tanto migliore sarà la stima di potenza prodotta dal modello. Per far fronte alle non-linearità e aumentare la significatività del processo di caratterizzazione PPP prevede l'utilizzo di modelli di regressione non-parametrici e di algoritmi di caratterizzazione adattativi.

I modelli non parametrici sono in grado di adattare al campione non solo il valore di alcuni coefficienti, ma la loro stessa struttura, rendendola più idonea a seguire eventuali non-linearità [10]. Gli algoritmi di caratterizzazione adattativi consentono di adattare il modello all'effettivo contesto operativo nel quale il circuito lavora [11].

4. Implementazione e risultati

La piattaforma di simulazione di PPP è *Verilog-XL*, un efficiente simulatore *event-driven* che garantisce la completa compatibilità con gli strumenti di progettazione basati sull'uso di *Verilog HDL*. I modelli di potenza sono stati implementati in C e *linkati* al simulatore attraverso la *Programming Language Interface* di *Verilog-XL*. Per validare i modelli e le strategie di simulazione sono stati utilizzati i *benchmark* dell'*iscas* e i moduli di un addizionatore *floating point* a doppia precisione.

Il simulatore logico di PPP produce stime di potenza globali e locali con errori rispetto a *HSPICE* inferiori al 5% e tempi di simulazione mille volte più veloci [7]. Il simulatore *behavioral* aumenta ulteriormente l'efficienza di due/tre ordini di grandezza con un'accuratezza media che varia dal 20% dei modelli lineari al 10% dei modelli adattativi [11].

PPP è divenuto il primo elemento di un ambiente integrato per sintesi e simulazione di circuiti CMOS a basso consumo di potenza. L'interfaccia utente di PPP è completamente basata sull'uso di tecnologie *Web*, che ne consentono l'uso diretto attraverso *Internet*. L'utente accede a PPP con un qualsiasi *Web browser* e lo utilizza remotamente grazie ad una serie di pagine HTML altamente interattive generate in modo dinamico durante la sessione di lavoro. Una versione prototipale di PPP è disponibile al seguente indirizzo: `http://akebono.stanford.edu/users/PPP`.

Bibliografia

[1] J. M. Rabaey and M. Pedram, *Low-power design methodologies*. Kluwer Academic Pub.s, 1996.

[2] F. Najm, "A survey of power estimation techniques in VLSI circuits," *IEEE Tran. on VLSI Systems*, vol. 2, no. 4, pp. 446–455, 1994.

[3] L. Benini *et al.*, "Analysis of hazard contribution to power dissipation in CMOS IC's," in *Proc. of IWLPD*, pp. 27–32, 1994.

[4] B. J. George *et al.*, "Power analysis and characterization for semi-custom desing," in *Proc. of IWLPD*, pp. 215–218, 1994.

[5] H. Sarin *et al.*, "A power modeling and characterization method for logic simulation," in *Proc. of IEEE Custom ICs Conf.*, pp. 363–366, 1995.

[6] J.-Y. Lin *et al.*, "A cell-based power estimation in CMOS combinational circuits," in *Proc. of IEEE ICCAD*, pp. 304–309, 1994.

[7] A. Bogliolo *et al.*, "Power Estimation of Cell-Based CMOS Circuits," in *Proc. of DAC*, pp. 433 – 438, 1996.

[8] P. Landman *et al.*, "Architectural power analysis, the Dual Bit Type method," *IEEE Tran. on VLSI Systems*, vol. 3, no. 2, pp. 173–187, 1995.

[9] H. Metha *et al.*, "Energy characterization based on clustering," in *Proc. of DAC*, pp. 702–707, 1996.

[10] L. Benini *et al.*, "Regression models for behavioral power estimation," in *Proc. of PATMOS*, 1996.

[11] A. Bogliolo *et al.*, "Adaptive least mean square behavioral power modeling," to appear in *Proc. of ED&TC*, 1997.