ESCLUSIVA

Come ottenere un compromesso ottimale tra dissipazione e tolleranza

Nilanjan Banerjee
graduate student in ECE at Purdue University

Kaushik Roy
Design Community Chair for the 45th Design Automation Conference (DAC) Executive Committee and the Roscoe H. George Professor of ECE at Purdue University in West Lafayette, Indiana

L'adozione di opportune tecniche progettuali permetterà di proseguire senza problemi di sorta sulla via tracciata dalla legge di Moore

La drastica riduzione delle geometrie dei dispositivi CMOS nell'ultimo trentennio ha permesso all'industria dei semiconduttori di soddisfare le crescenti esigenze in termini di capacità di calcolo e densità d'integrazione. La tecnologia CMOS, nonostante abbia soddisfatto positivamente le richieste provenienti dal mondo industriale, sta incontrando ostacoli di notevole entità nei nodi tecnologici inferiori a 65 nm, imputabili ai livelli di integrazione ancora maggiori e alle limitazioni di natura fisica dei dispositivi. Dissipazione di potenza e variazioni di processo sono i problemi più critici che i progettisti di integrati devono affrontare oggigiorno.
I componenti fondamentali della dissipazione di potenza sono la potenza dinamica e la potenza statica. La prima è aumentata significativamente nel corso dell'evoluzione delle diverse generazioni tecnologiche a causa delle sempre maggiori frequenze di clock e della commutazione simultanea di un elevato numero di transistor su un singolo die. La miniaturizzazione dei transistor si è tradotta in un incremento delle correnti di perdita e in una più elevata dissipazione statica. A causa della limitata capacità di raffreddamento dei package, una maggiore dissipazione di potenza all'interno della minuscola area del die genera temperature elevate e provoca l'insorgere di problemi di affidabilità di notevole entità. Nelle applicazioni dove è prevista l'alimentazione a batteria, il maggiore

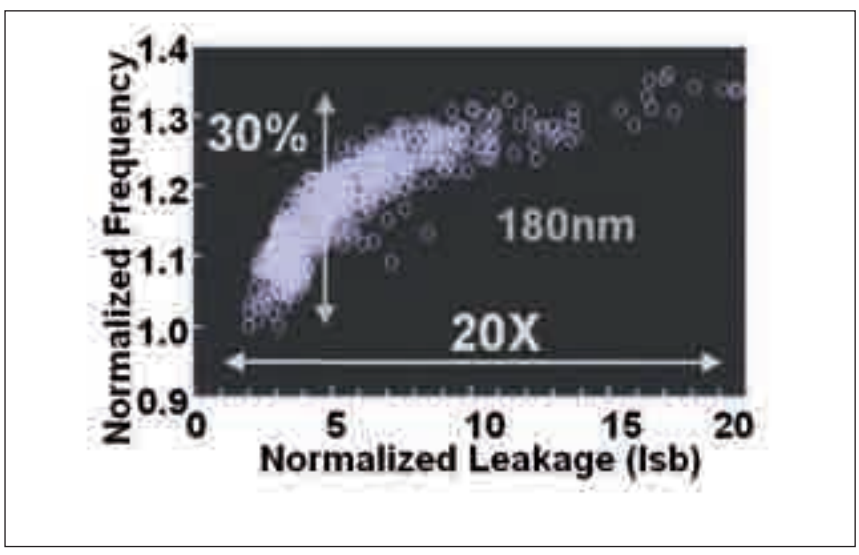


Fig. 1 – Variabilità per transistor da 180 nm (fonte: Intel)

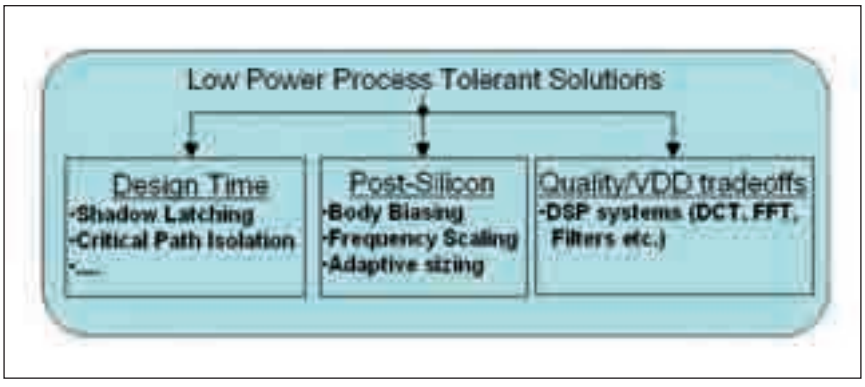


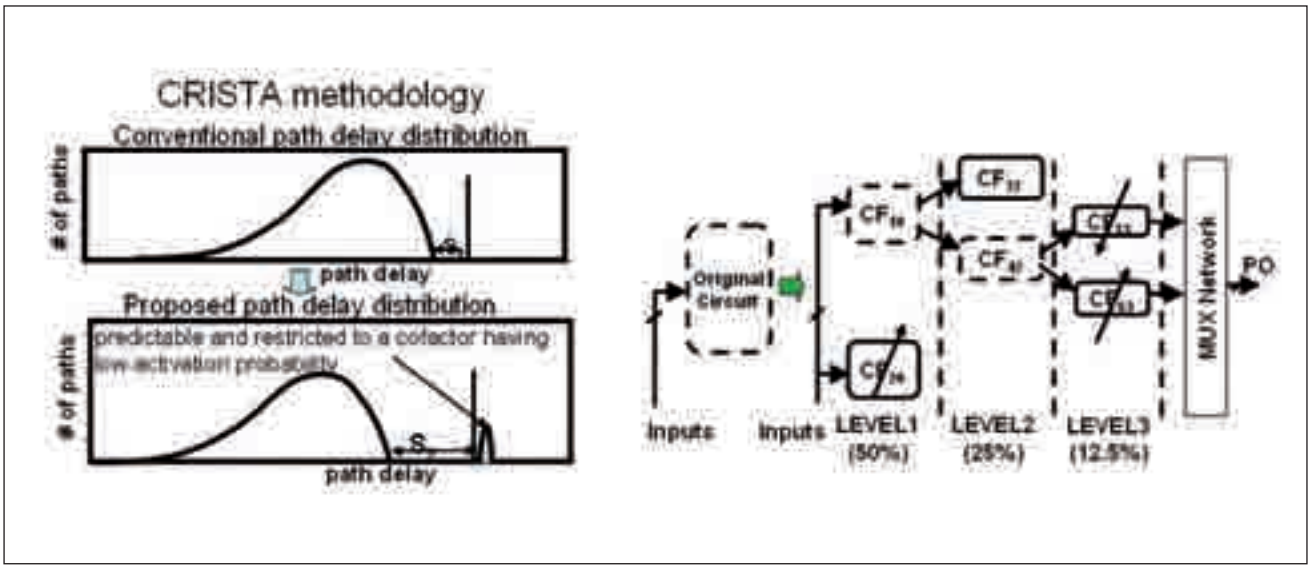
Fig. 2 – Soluzioni che permettono di ottenere un compromesso tra tolleranza di processo/bassa potenza

consumo di potenza comporta una sensibile diminuzione della durata delle batterie. Appare dunque evidente che il problema della potenza riveste una notevole importanza nel progetto degli odierni circuiti integrati.

Le variazioni di processo, d'altra parte, sono imputabili ai limiti intrinseci dei processi di fabbricazione (come la litografia al di sotto della lunghezza d'onda, i meccanismi di incisione e drogaggio, e così via). Tali limiti si manifestano sotto forma di variazioni nei parametri fisici come lunghezza, larghezza, spessore dell'ossido, e così via. Per le tecnologie al di sotto dei 65 nm, il numero ed il posizionamento degli atomi di drogaggio nel canale possono causare notevoli variazioni fra transistor differenti. Le variazioni dei parametri di processo possono verificarsi a diversi livelli di granularità: all'interno del die, fra die e die e fra wafer e wafer. Le discrepanze nei parametri del dispositivo, di natura sistematica oppure casuale, si traducono in variazioni dei parametri circuitali come ad esempio il ritardo, che comportano perdite nella resa parametrica che si manifestano attraverso variazioni dei tempi di ritardo (Fig. 1). Per mantenere una resa adeguata nei progetti che prevedono l'uso delle tecnologie odierne e di quelle future, è importante affrontare anche il 'problema delle variazioni di processo'. Nonostante le tensioni di alimentazione più elevate e le maggiori dimensioni dei transistor contribuiscano ad attenuare la portata di alcuni di questi problemi, essi pongono requisiti contrastanti in termini di dissipazione di bassa potenza.

Le esigenze in termini di bassa dissipazione e tolleranza alle variazioni di processo, in contrasto tra di loro, rendono senza dubbio più difficile la progettazione degli odierni circuiti integrati. Per affrontare con efficacia queste problematiche nel momento in cui verranno adottate le future tecnologie è auspicabile l'apporto di innovazioni a livello di sistema. Per ridurre la potenza dinamica e statica negli attuali progetti di inte-

Fig. 3 – CRISTA: una tecnica di progettazione basata sui tempi del progetto per sviluppare sistemi a bassa potenza tolleranti alle variazioni



Progetti sempre più efficienti dal punto di vista energetico

I progetti in cui il fattore energetico viene tenuto nella debita considerazione sono indispensabili per prolungare la durata delle batterie nei dispositivi portatili, evitare l'eccessiva generazione di calore che potrebbe comportare problemi di affidabilità dei dispositivi e ridurre i costi associati all'adozione di costose tecniche di raffreddamento. Per fare fronte alla crescente domanda di messaggistica video nelle comunicazioni multimediali/wireless, è necessario sviluppare schemi di trasmissione di immagini/video a bassa energia. I tradizionali schemi di compressione delle immagini sono stati progettati per minimizzare la distorsione dell'immagine ricostruita per una data velocità di trasmissione. Le applicazioni multimediali utilizzate nelle apparecchiature portatili potrebbero non richiedere sempre la massima qualità dell'immagine. Questo aspetto può essere sfruttato per ottenere architetture in grado di offrire il corretto compromesso fra qualità dell'immagine e consumo di energia. La DCT (Trasformata cosinusoidale discreta) risulta particolarmente utile nel settore della compressione video e di immagini grazie alla sua capacità intrinseca di garantire elevati rapporti di compressione a fronte di una bassa complessità progettuale. Molte ricerche sono state condotte per ridurre il numero e la complessità dei calcoli per le architetture DCT a bassa potenza. La minimizzazione della potenza non è il solo requisito dei progetti attuali. A causa della progressiva riduzione delle geometrie, anche le variazioni dei parametri di processo rappresentano un importante problema progettuale. È stato infatti dimostrato che le variazioni parametriche creano una dispersione dei ritardi pari a circa il 30% per la tecnologia di processo da 70 nm, il che comporta l'insorgere di

grati sono state proposte numerose tecniche. A causa della dipendenza quadratica della potenza dinamica di un circuito dalla sua tensione di funzionamento, la progressiva diminuzione della tensione di alimentazione si è dimostrata un metodo estremamente efficace nella riduzione della dissipazione di potenza. Non va comunque dimenticato che la riduzione della tensione di alimentazione si traduce in maggiori ritardi sui percorsi e rende i progetti più inclini a errori di ritardo. Per migliorare il rapporto tra potenza e area sono state adottate altre procedure di ottimizzazione, tra cui il dimensionamento dei gate, la doppia assegnazione di Vt e la doppia alimentazione. Benché efficaci per quel che concerne la limitazione dei consumi di potenza, tali metodi contribuiscono a un aumento del numero di percorsi critici in un circuito, rendendolo quindi più vulnerabile a fenomeni di ritardo nel caso di variazioni dei parametri. Per cercare di contrastare le elevate perdite che si verificano quando si adottano tecnologie su scala nanometrica, sono stati sviluppati anche numerosi schemi di riduzione delle perdite.

L'approccio più comune è l'uso di "transistor impilati" (stacking transistor). Benché questa tecnica riduca considerevolmente la dissipazione di potenza statica, essa non ha effetto alcuno sugli effetti nocivi delle variazioni dei parametri. D'altra parte, per minimizzare il numero di guasti sui chip dovuti alle variazioni, si può adottare un approccio di progettazione 'conservativo', assicurando cioè il funzionamento del circuito all'interno delle specifiche previste anche nel caso si verifichi il "caso peggiore" (worst case). Ciò può essere ottenuto mediante lo "scaling" VDD o l'aumento delle dimensioni dei gate logici. Entrambe queste metodologie contribuiscono a ridurre le divergenze di processo a scapito di un significativo "appesantimento" in termini di potenza, prestazioni e occupazione di area. Per evitare queste problematiche, sono state di recente introdotte metodologie di progettazione statistica.

Esse prevedono la modellazione di alcuni parametri circuitali (per esempio ritardo o perdita) sotto forma di distribuzioni statistica (ad esempio gaussiana) e la progettazione circuitale è finalizzata al soddisfacimento di un vincolo di resa in funzione di un valore "target" del parametro.

Nella maggior parte dei casi, queste tecniche prevengono la progettazione nel "caso peggiore" ma non sono in grado di assicurare una contemporanea riduzione dei consumi.

Quando detto in precedenza dimostra che le tolleranze di processo e la bassa potenza sono requisiti di progettazione in conflitto tra loro. A questo punto è necessario un cambiamento radicale di approccio alla progettazione di sistema: in particolare, occorre sviluppare nuove soluzioni a livello di sistema per risolvere il contraddittorio bassa potenza/tolleranze di processo.

Di recente sono state avviate ricerche di soluzioni per il problema della bassa potenza, che prevedono una drastica riduzione della tensione di alimentazione nel rispetto dei requisiti di resa temporale. Queste tecniche possono essere grosso modo suddivise in tre categorie principali: a) temporizzazioni presenti nel progetto, b) post-silicio e

problemi di ritardo in alcuni chip. Il buon senso suggerisce un approccio progettuale conservativo (per es. l'aumento della Vdd o il sovradimensionamento dei gate logici) per evitare i problemi di ritardo ed ottenere un'elevata resa parametrica. Tuttavia, queste tecniche comportano un aumento della potenza e/o degli ingombri sul die. È indubbio che tolleranza di processo e bassa potenza rappresentano requisiti progettuali in conflitto tra loro. La metodologia proposta permette di affrontare simultaneamente le problematiche legate alla bassa dissipazione e alla tolleranza di processo grazie all'adozione di un'architettura in grado di supportare la riduzione della tensione anche in presenza di variazioni parametriche. Le fasi salienti si possono così riassumere:

- identificazione dei percorsi di calcolo indispensabili per mantenere un'elevata qualità dell'immagine;

- sviluppo di un algoritmo/un'architettura che sia in grado di garantire i percorsi più brevi per i calcoli più importanti (in termini di qualità dell'immagine) rispetto a quelli dei percorsi meno importanti;

- utilizzo di questa architettura per garantire la prevedibilità di eventuali errori di ritardo sui percorsi in presenza di una tensione di alimentazione singola ridotta e di variazioni dei parametri di processo e possibilità di tollerare gli errori in tali percorsi con una degradazione minima sul PSNR dell'immagine;

- riconfigurazione dell'architettura in modo da ottenere un corretto compromesso tra qualità dell'immagine e consumo di potenza.

I risultati dimostrano che, anche nel caso di grandi variazioni del processo e della tensione di alimentazione (0,8V), l'architettura proposta evidenzia una degradazione graduale della qualità dell'immagine a fronte di considerevoli risparmi di potenza (63%) rispetto alle attuali implementazioni nella tecnologia di processo da 70 nm.

c) sistemi che prevedono la possibilità di compromessi tra qualità e i requisiti di potenza (Fig. 2).

Tecniche basate sulla temporizzazione

Uno schema molto diffuso che rientra in questa categoria è chiamato Razor [1], basato sul concetto di riduzione dinamica della tensione (DVS), dove è possibile ottimizzare i risparmi di potenza mediante una drastica riduzione della tensione di alimentazione, pur assicurando il corretto funzionamento del progetto. Il meccanismo DVS previsto dal metodo Razor si basa sul rilevamento e sulla correzione dinamica degli errori di temporizzazione nel circuito. L'idea chiave è quella di regolare la tensione di alimentazione monitorando il tasso di errori durante il funzionamento del circuito, eliminando in tal modo la necessità di margini di tensione e sfruttando la dipendenza dai dati dei ritardi circuitali. Viene introdotto un flip-flop custom che esegue un doppio campionamento degli stadi del pipeline, una volta con un clock veloce e successivamente con un clock ritardato. Nel caso di un errore di temporizzazione (che di verifica nel momento in cui i valori campionati non coincidono), un meccanismo modificato di recupero degli errori di specularità sulla pipeline ripristina lo stato corretto del programma nel progetto. Un metodo recentemente proposto [2] è riuscito a sintetizzare con esito positivo progetti a bassa potenza tolleranti alle variazioni mediante l'isolamento dei percorsi critici nei circuiti combinatori. La nozione di isolamento dei percorsi critici si riferisce al confinamento dei percorsi critici in un blocco logico noto all'interno del progetto. Questo approccio porta all'adozione di una metodologia di progettazione che prevede la possibilità di isolare e prevedere i gruppi di percorsi che potrebbero diventare critici nel caso di variazioni parametriche. Ogni possibile malfunzionamento imputabile al ritardo in questi percorsi viene gestito attraverso un'opportuna modifica di natura adattativi del clock. A condizione che questi percorsi vengano attivati raramente, viene garantito un "overhead" veramente minimo sulle prestazioni.

Aumentando il margine di ritardo fra i percorsi critici e non critici tramite la sintesi logica e l'opportuno dimensionamento dei gate, è possibile ottenere una resa maggiore in presenza di variazioni parametriche e una sensibile riduzione della tensione per diminuire la potenza. Questa tecnica può essere utilizzata sia a circuiti di logica sparsa sia a unità di esecuzione.

Tecniche post-silicio

Tali metodologie prevedono il rilevamento delle condizioni di processo (process corner) nelle quali si trova il die (corner process) e il ricorso a misure adattative per migliorare la resa parametrica. La condizione di processo nella quale si trova quale un chip può essere determinata mediante la rilevazione di perdite o ritardi.

Le maggiori perdite/i minori ritardi si verificano nella condizione in cui Vt ha un valore basso, mentre le minori perdite e i maggiori ritardi sono normalmente associati a un valore di Vt elevato. Fra le tecniche di progettazione post-silicio, quella della polarizzazione adattativa del corpo si è dimostrata estremamente

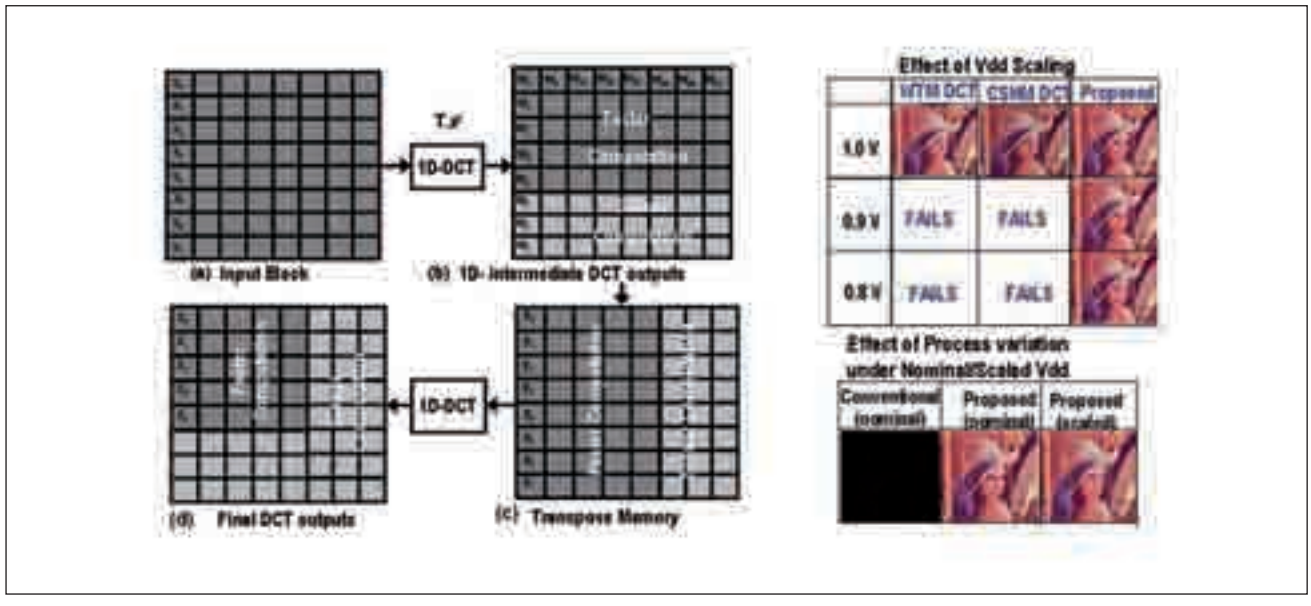


Fig. 4 – a) Concetto di DCT tollerante al processo b) Effetti dello scaling di Vdd e delle variazioni sulle uscite dei progetti che utilizzano tecniche convenzionali e quelli che adottano la tecnologia proposta nell'articolo

efficace nella minimizzazione delle perdite di potenza e nella compensazione di diverse variazioni di processo. Il punto di partenza per una polarizzazione ottimale del corpo è rappresentato dal rilevamento della condizione di processo in cui si trova il blocco circuitale. Tale informazione viene quindi utilizzata per determinare l'ampiezza e la polarità della tensione di polarizzazione del substrato ottimizzata in modo da realizzare progetti a basse perdite che risultino 'consapevoli delle variazioni'. Questa tecnica è stata utilizzata per le logiche e le memorie al fine di migliorare la resa parametrica in presenza di variazioni sia sistematiche sia casuali.
Il dimensionamento adattativo (per es. l'uso di transistor di mantenimento programmabili per logica tipo domino) è stato anche utilizzato per lo sviluppo di progetti "affidabili" di file di registri ad alte prestazioni.

Compromesso tra qualità e Vdd

Per un gran numero di sistemi di elaborazione dei segnali a bassa potenza, i progettisti usano un altro parametro, quello della 'qualità' per ottenere un equo compromesso per quel che concerne la dissipazione di potenza. Infatti, un attento bilanciamento tra la qualità e i requisiti di potenza può portare a una riduzione della tensione, con conseguente diminuzione della dissipazione, anche in presenza di variazioni dei parametri di processo. Di conseguenza, l'adozione di idonei algoritmi e architetture può portare a una diminuzione accettabile della 'qualità' abbinata a una riduzione della tensione di alimentazione.
L'architettura DCT (Trasformata cosinusoidale discreta), ad esempio, permette di ottenere una drastica riduzione della tensione in virtù della considerazione che non tutti i calcoli intermedi sono ugualmente importanti in un sistema DCT per ottenere una 'buona' qualità dell'immagine con un rapporto segnale di picco/rumore (PSNR) > 30 dB. Sulla base di questa osservazione, i percorsi di segnale della DCT che contribuiscono in misura inferiore al miglioramento del PSRN sono progettati in modo da risultare più lunghi rispetto ai percorsi che contribuiscono in misura maggiore al miglioramento di questo parametro.
In presenza di riduzioni della tensione di alimentazione e/o variazioni dei parametri di processo, tutti i possibili errori di ritardo si manifestano solamente sui percorsi lunghi, che hanno cioè un peso inferiore rispetto per quel che riguarda il miglioramento del PSNR: in questo modo risulta possibile ottenere notevoli migliorie in termini di dissipazione di potenza a fronte di un deterioramento di entità trascurabile del PSNR.
In definitiva, i metodi discussi indicano una tendenza in atto nel settore della progettazione di integrati, dove le innovazioni a livello di sistema permetteranno di risolvere contemporaneamente sia i problemi di riduzione della potenza sia quelli relativi alla tecnologia di processo.
I continui sforzi di ricerca finalizzati allo sviluppo di progetti a bassa potenza e in grado di supportare le variazioni di processo permetteranno di proseguire senza problemi sulla via tracciata dalla legge di Moore.