MRAM per il cognitive computing

La tecnologia MRAM di Everspin è alla base delle soluzioni di memoria CogniBlox di CogniMem Technologies che aprono la strada a maggiori potenzialità di elaborazione per numerose applicazioni

Francesco Ferrari

Le memorie MRAM, acronimo di Magnetic RAM, di Everspin sono i componenti principali per la nuova piattaforma di computing cognitivo di CogniMem Technologies. La piattaforma si basa sui moduli riconfigurabili che utilizzano i chip CM1K . Questi chip permettono di implementare 4096 neuroni di silicio che imitano le modalità con cui quelli umani elaborano le informazioni.
Ogni modulo CogniBlox utilizza 4 MB di MRAM Everspin per memorizzare i knowledge file (modelli appresi) ma anche immagini video o altro. La scheda può essere riconfigurata ricorrendo a un FPGA.
La memoria MRAM ha la peculiarità di offrire una velocità paragonabile a quella delle memorie SRAM (l'ordine di grandezza è quello dei nanosecondi), compresa la capacità di non perdere le informazioni memorizzate nel caso venga tolta l'alimentazione; caratteristiche che ne rendono idoneo l'impiego per applicazioni di pattern recognition utilizzate per il cognitive computing.
Le informazioni nelle MRAM sono memorizzate nel materiale magnetico integrato nel silicio del circuito del chip.
I CogniBlox sono moduli stackable e riconfigurabili che offrono versatilità nella progettazione dei sistemi, una architettura per il riconoscimento dei pattern necessario alle applicazioni di cognitive computing, ma anche all'analisi video e altre applicazioni che spaziano, per esempio, dalle funzioni di visione per le macchine al riconoscimento vocale, al data mining oppure per sistemi di sorveglianza real time, previsioni merceologiche e un ampio numero di elaborazioni scientifiche.
A differenza delle tradizionali tecniche di Von Neumann che devono fare i conti con problematiche come le prestazioni delle memorie di CPU e GPU, la sincronizzazione e le comunicazioni a causa dell'accesso alla memoria di tipo seriale, CogniBlox ha il vantaggio di elaborare e accedere alla memoria in parallelo.
Un altro utilizzo delle memorie MRAM nella piattaforma CogniBlox deriva dalla capacità di salvare e ripristinare i knowledge file nei neuroni in silicio quando sta funzionando un'applicazione.
L'esempio che viene fatto dal produttore è inerente le informazioni che, memorizzate nei neuroni, permettono di riconoscere una sagoma di una persona e, se viene identificata una presenza umana nel campo visivo, il contenuto dei neuroni può essere modificato per memorizzare il secondo gruppo di informazioni e provvedere all'identificazione.

Per quanto riguarda le potenzialità di questa tecnologia, CogniBlox permette di configurare un sistema con un milione di neuroni in silicio utilizzando 250 schede che possono raggiungere le performance di 0,13 petaops con un consumo tipico di 500 W e la possibilità di realizzare 256 milioni di connessioni ogni 10µs. Il sistema CogniBlox è supportato sotto le piattaforme Windows e Linux utilizzando i tool .NET e Java per i comandi di training e recognition dell'array. ■

L'ottimizzazione del ciclo produttivo è più efficace con una innovativa tecnica MLR

Una tecnica innovativa consente di migliorare il rendimento del processo di fabbricazione dei chip sui wafer tenendo conto della variabilità dello spessore del die

Philippe Morey - Chaisemartin

Eric Beisser - Xyalis

Alcune fasi nel ciclo di fabbricazione dei chip sui wafer causano spesso la non trascurabile comparsa di varianti non congruenti con il progetto originale dei chip in produzione. Tipicamente si tratta di eventi occasionati durante la planarizzazione chimica e meccanica, Chemo-Mechanical Planarization, CMP, che introducono variazioni non omogenee dello spessore del wafer che rimangono irregolari per l'intero corso del processo di assemblaggio. Queste variazioni hanno un impatto sulle prestazioni del chip e quindi anche sulla fase finale di ottimizzazione con le tecniche MLR, Multi Layer Reticles. Inoltre,

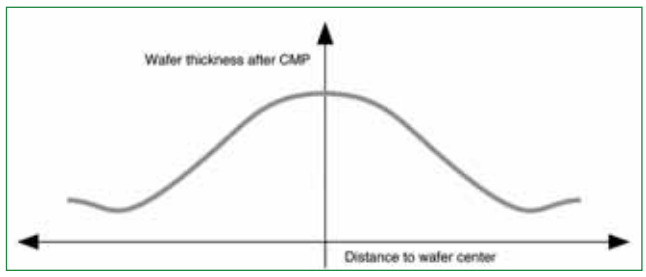

Fig. 1 – Variazione media dello spessore in un wafer

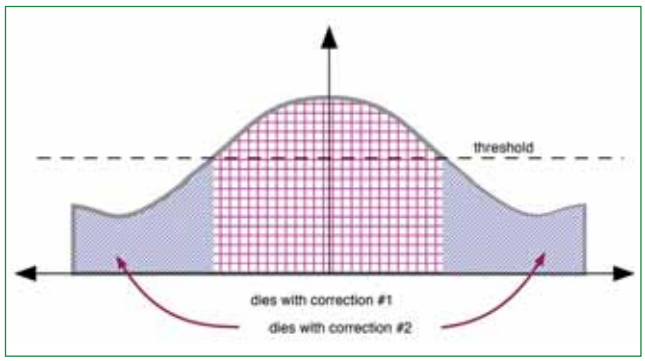

Fig. 2 – Suddivisione di due regioni di wafer con diverso spessore medio determinate dall'applicazione di una soglia

questo effetto può essere ulteriormente amplificato allorquando le variazioni di spessore riescono a creare dei dislivelli capaci di indurre in errore la successiva fase di fotolitografia.
D'altra parte, il layout originale dei chip può essere modificato inserendo delle correzioni capaci di compensare queste variazioni di spessore, ma si tratta di correzioni determinate da modelli statistici del wafer che possono rimediare solo ai valori medi degli errori calcolati considerando l'intero wafer. Per risolvere adeguatamente questa problematica è, invece, meglio una metodologia capace di correggere i singoli errori mediando fra i soli valori adiacenti come si fa nelle tecniche di ottimizzazione MLR. In pratica, cambia il punto di vista perché le maschere non sono più la composizione di diverse rappresentazioni riguardanti l'intero chip, ma sono disegnate utilizzando più immagini ottimizzate di un'unica rappresentazione del chip.

Le problematiche

In tutti i metodi classici per l'ottimizzazione del rendimento della fase CMP il layout viene corretto inserendo opportune forme fittizie nel progetto originale che vengono decise in base a stime statistiche sulle variazioni medie dello spessore del die. Molti recenti studi hanno dimostrato che queste stime sulle variazioni non descrivono adeguatamente ciò che realmente succede nel die, mentre è accertato che le variazioni di spessore possono avere un forte impatto sulle prestazioni dei chip e contribuire a produrre chip non funzionali o con prestazioni diverse da quelle previste. Fortunatamente, le irregolarità del processo nella fase CMP possono essere previste con una tecnologia innovativa che consente di tracciare l'andamento dello spessore del wafer a partire dal centro del die via via fino all'esterno dei chip realizzati sopra a esso. Tenendo conto di questo parametro aggiuntivo oltre che delle variazioni medie stimate statisticamente è possibile migliorare il rendimento globale della fase CMP e anche delle successive fasi di ottimizzazione sul layout dei chip. Si tratta, dunque, di una fase di ottimizzazione DFM (Design For Manufacturability) che consente di migliorare il posizionamento delle forme fittizie di correzione sul layout originale tenendo conto degli errori medi stimati e anche degli errori localizzati valutando la variazione dello spessore del wafer dal centro del die. Come risultato si ottiene un miglior livellamento dello spessore del wafer e, inoltre, si riesce a tenere conto delle variazioni di spessore evidenziate nella modellizzazione OPC (Optical Proximity Correction). Con questa metodologia si riescono a correggere molti degli errori tipici introdotti nella fase CMP ed eliminarne gli effetti sulle prestazioni dei chip fabbricati. Il costo di questa tecnologia è competitivo perché non richiede attrezzature extra, ma solo qualche economico adattamento delle tecniche di ottimizzazione attualmente già molto diffuse (MLR, CMP e OPC).

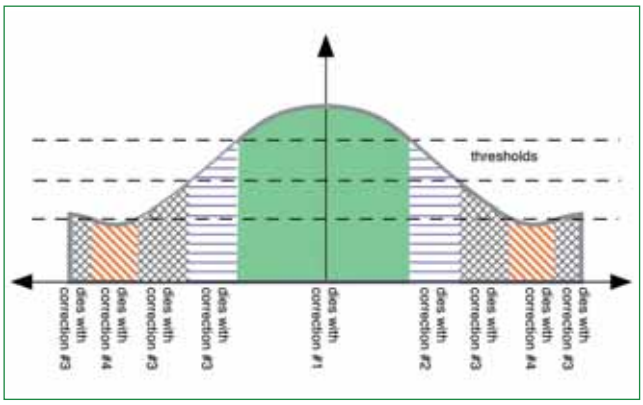

Fig. 3 – Applicando tre livelli di soglia sullo spessore del wafer si ottengono cinque regioni descritte da altrettanti valori di spessore medio localmente più accurati

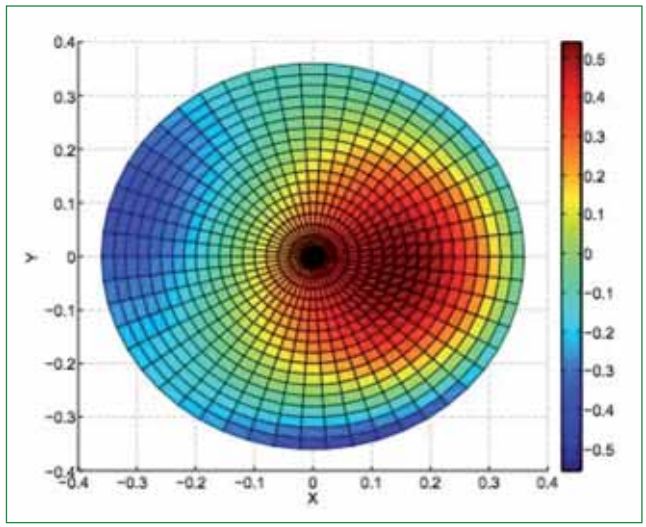

Fig. 4 – Esempio di distribuzione non simmetrica della variazione dello spessore in un die

La metodologia

È indubbio che nei processi di fabbricazione dei wafer possono introdursi variazioni irregolari dello spessore del die. Un andamento medio della variazione dello spessore nella fase CMP può essere stimato nella forma gaussiana che si vede nella prima figura. Nel modo di procedere classico a questo punto si determinano le forme fittizie di correzione da applicare

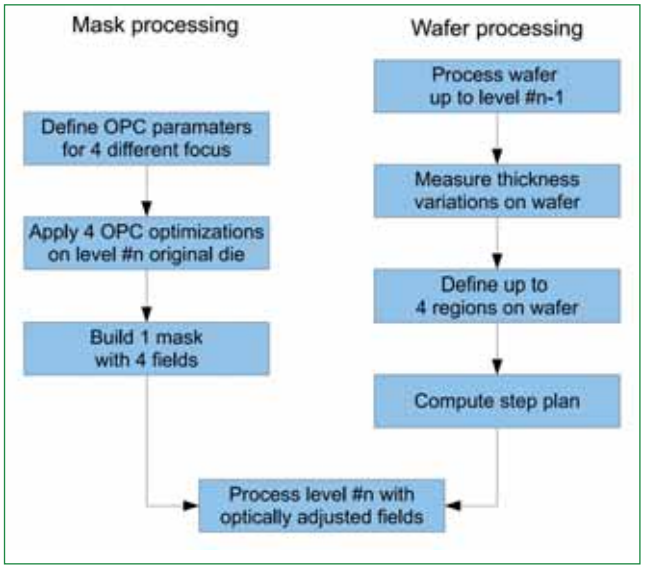

Fig. 5 – Schema a blocchi di un esempio di procedura di ottimizzazione che consente di tenere conto del modello OPC

per compensare le variazioni di spessore introdotte nella fase CMP e per far ciò ci si basa sull'andamento medio dei valori di spessore dell'intero wafer. Il posizionamento e l'inserimento di queste correzioni nel modello originale comportano un notevole dispendio di tempo e, inoltre, nel modello si tiene conto solo di alcuni parametri generici come la densità dei componenti o le caratteristiche dei materiali presenti. La novità della tecnica ideata dagli esperti Xyalis consiste nell'aggiungere il nuovo parametro costituito dalla distanza di ogni punto del wafer dal suo centro geometrico. Si tratta, in pratica, di definire alcune regioni del wafer dove si riesce a ottenere una stima localmente più precisa dello spessore del wafer. Ciascuna regione del wafer viene, dunque, descritta con un parametro di misura dello spessore localmente più preciso. Nella seconda figura si vede un esempio dove ci sono due regioni del wafer determinate dall'applicazione di un unico livello di soglia, ma se i livelli di soglia sono tre allora si ottengono cinque regioni del wafer con altrettanti cinque valori medi di spessore come si vede nella terza figura. Quanti più livelli di soglia si definiscono e tanto più accurato sarà il modello del die per la fase CMP. Nella metodologia classica di ottimizzazione occorre prevedere più maschere di fotolitografia perché devono adattarsi allo spessore medio di ciascuna regione. La tecnica comunemente nota come MLR, Multi Layer Reticles, prescrive di suddividere il reticolo della maschera principale per la fotolitografia in più regioni che rappresentano altrettanti livelli di spessore medio nel die e generalmente non sono mai più di quattro o cinque. Naturalmente c'è il rischio che aumentando il numero delle regioni si complichi in ugual misura il processo di fotolitografia e si moltiplichi il costo delle maschere e occorre, dunque, cercare di evitare che ciò succeda.

La novità introdotta dagli esperti Xyalis consiste nell'utilizzare questa tecnica non per definire diversi livelli di spessore medio nel reticolo, ma per creare diverse descrizioni dello spessore riguardanti un unico livello del reticolo. In questo modo si può tenere conto delle variazioni di spessore tipiche di ciascuna regione e migliorare l'efficacia del posizionamento delle correzioni fittizie mentre, nel contempo, si continua a utilizzare una sola maschera e perciò non s'innalzano i costi. L'unica complessità in più che costituisce perciò anche un costo aggiuntivo si deve ai diversi tempi di esposizione da prevedere nella fotolitografia delle singole regioni. In ciascuna regione il laser deve, infatti, ridurre o prolungare l'esposizione per frazioni di 1/2 o 1/4 rispetto al tempo standard in modo tale da tenere conto della minore o maggiore distanza dalla superficie del wafer rispettivamente dovuta al maggiore o minore spessore del die. Con questo metodo si possono impostare anche punti specifici nel wafer dove occorrono condizioni di esposizione particolari dovute a variazioni di spessore singolari. Per esempio, si può tenere conto delle singolarità identificate dal modello OPC e prevedere condizioni di illuminazione laser singole da utilizzare solamente in quei punti del wafer.Inoltre, se la variazione dello spessore non risulta simmetrica come capita nel caso generico, se ne può tenere conto modificando i tempi di esposizione in modo da prenderla in considerazione, ad esempio, nella quarta figura. Si può anche prevedere di fare qualche misura preliminare su ogni wafer all'inizio della lavorazione e ciò può probabilmente comportare dei costi in più che però tendono presumibilmente a diminuire man mano che si diventa più esperti nell'impiego di questa tecnica di ottimizzazione.

I costi dell'ottimizzazione

Per valutare i costi globali di questa metodologia occorre tenere conto sia dei costi che si aggiungono nella fabbricazione dei chip sia del miglioramento che si consegue grazie all'ottimizzazione del processo. I costi di fabbricazione aggiuntivi sono dovuti alle più piccole dimensioni delle regioni definite come MLR ed è indispensabile, comunque, fare un'analisi preliminare dettagliata per valutare se c'è maggior convenienza nei costi con una suddivisione del wafer in due o in quattro regioni.

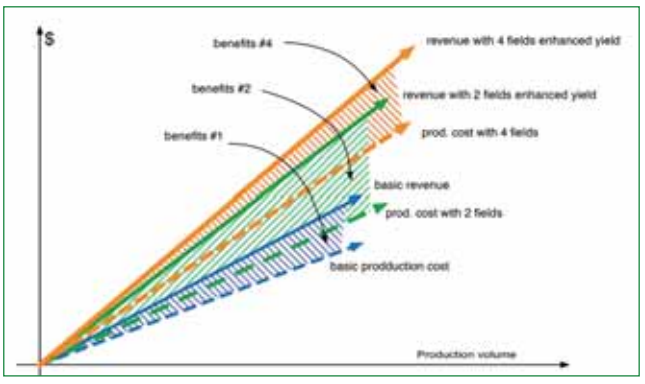

Fig. 6 – Esempio di analisi dei vantaggi ottenibili nei costi con una suddivisione del die in due o quattro regioni

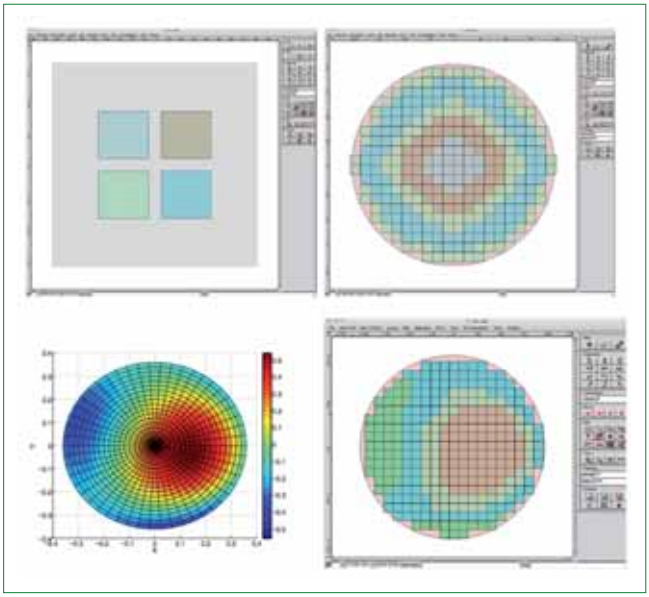

Fig. 7 – Pianificazione della fotolitografia con quattro livelli di soglia (sopra) e un esempio di pianificazione dinamica che tiene conto della variazione asimmetrica dello spessore medio del die (sotto)

Il criterio di scelta fra nessuna, una, due, tre o quattro suddivisioni del die dipende soprattutto dalle variazioni di processo che si prevedono a livello del wafer e all'impatto che ci si aspetta che esse abbiano sulle prestazioni dei chip. Si consideri, ad esempio, che una brusca variazione di spessore in prossimità della giunzione di un transistor può portare a conseguenze drammatiche sul funzionamento dell'intero circuito. D'altra parte è ben noto a tutti che la fotolitografia è la fase più critica del costo di produzione di un wafer. Dunque, i profitti che si otterranno nelle vendite dei chip dipendono dai costi di produzione che sono a loro volta legati all'efficienza del processo di fabbricazione e quindi alle prestazioni della fase di fotolitografia. È chiaro perciò che la metodologia descritta consente di migliorare considerevolmente questa fase del processo di fabbricazione e quindi ridurre in ugual misura i costi globali.

La pianificazione del processo

Il criterio di scelta adottato nell'ottimizzazione di un wafer ne determina la qualità finale dato che ciascuna delle regioni in cui è suddiviso può essere lavorata con miglior rendimento. Infatti, il wafer viene descritto da una mappa di regioni di die con diverso spessore medio e utilizzando tempi di esposizione appropriati per ciascuna regione, si può calibrare più accuratamente il processo di fotolitografia. È necessario, tuttavia, preparare un'opportuna pianificazione o "step plan" che consente di assicurare la corretta lavorazione su ciascuna regione del die. La pianificazione dei parametri di fotolitografia può essere generata automaticamente in funzione delle soglie di spessore decise in base all'analisi preventiva sulle variazioni di spessore stimate nel processo; si può anche decidere di correggere la pianificazione modificando le regioni durante il processo stesso e, quindi, ottenere una regolazione dinamica della fase di fotolitografia. In entrambi i casi questa fase diventa un po' più lunga perché impone la ricalibrazione dei parametri di esposizione e di focalizzazione del fascio laser su ciascuna delle regioni, ma la maggior durata del processo è compensata dall'importantissimo vantaggio di poter usare una sola maschera di fotolitografia e ciò consente di ridurre considerevolmente i costi.

La metodologia Xyalis consente di migliorare il rendimento del processo di fabbricazione dei wafer soprattutto durante la fase di fotolitografia perché consente di tenere conto delle variazioni di spessore del die pur continuando a utilizzare una sola maschera. Si può, pertanto, migliorare l'efficacia delle correzioni al layout utilizzando modelli noti e tecnologie già diffuse come CMP, OPC e MLR senza bisogno di investire in attrezzature specifiche. ■