

LA REGOLA DI RENT NELLE ARCHITETTURE FPGA

(I parte)

Catello Antonio de Rosa

Pubblicata nel 1960, la regola stabilisce un criterio di analisi all'interno di un circuito integrato che pone in relazione il numero delle porte logiche contenute in un blocco al numero di connessioni che collegano il blocco agli altri blocchi. In questa prima parte viene descritta la regola di Rent e la sua applicazione negli FPGA

La legge di Moore predice che ogni 18 mesi la capacità elaborativa di un circuito integrato raddoppia a parità di area. Questo si ottiene con uno shrink tecnologico di processo, in cui a dimensioni di canale sempre più ridotte corrisponde un maggior numero di transistor e quindi di interconnessioni, e con un aumento del numero dei livelli di metallizzazione. Questa previsione, enunciata nel 1965, resta ancora valida con sorpresa dello stesso Moore. Poiché ogni circuito integrato ha una dimensione limitata della sua periferia, il numero di segnali di I/O non può crescere in maniera simile a quello delle porte logiche interne. Ciò causa un "communication bottleneck" predetto dalla regola di Rent.

La legge di Rent

La regola di Rent è dovuta all'omonimo ingegnere dell'IBM che, nel 1960, pubblicò due relazioni tecniche interne. Questa regola empirica, avvalorata da successive sperimentazioni, stabilisce un criterio di analisi all'interno di un circuito integrato che pone in relazione il

numero delle porte logiche contenute in un blocco al numero di connessioni che collegano il blocco agli altri blocchi.

Le connessioni, o terminali, che delimitano l'attraversamento dei confini del blocco vengono generalmente definiti "pin". Le porte logiche, chiamate spesso "gate", sono transistor o gruppi di transistor opportunamente collegati, macrocelle o celle logiche di tipo semplice o composto. Una o più porte logiche formano un blocco.

La regola di Rent è basata sull'osservazione di circuiti integrati già esistenti e di architettura tradizionale; tuttavia può fornire un metodo di partenza per prevedere il comportamento di architetture di circuiti integrati di tipo non tradizionale. In pratica è possibile valutare la bontà di una nuova architettura ancor prima che questa sia fisicamente realizzata. Nel corso di questi ultimi quarant'anni è stata usata per studiare nuove architetture di computer o di chip, stimare parametri di layout, generare circuiti di benchmark, ed è risultata talmente affidabile da meritarsi l'appellativo improprio di legge di Rent.

Formulazione della regola di Rent

Se consideriamo una regione omogenea all'interno di un circuito integrato, ovvero un'area nella quale la lunghezza media delle interconnessioni per ciascun gate di logica e il numero medio di terminali per gate è indipendente dalla posizione all'interno della regione, la regola di Rent fornisce uno strumento per stimare il numero di connessioni che attraversano i confini della regione. Tali connessioni sono quelle necessarie per porre in comunicazione un blocco logico, cioè la regione, con il resto del sistema.

Si supponga di introdurre una perturbazione nella geometria della regione, ovvero si aggiunga un altro piccolo gruppo di porte logiche al blocco considerato (Fig. 1). La nuova regione definisce un confine superiore al precedente che sarà attraversato da un numero maggiore di connessioni. In assenza di altre informazioni è solo possibile stimare il numero addizionale di terminali necessari assumendo che per i gate aggiunti debba essere mantenuto lo

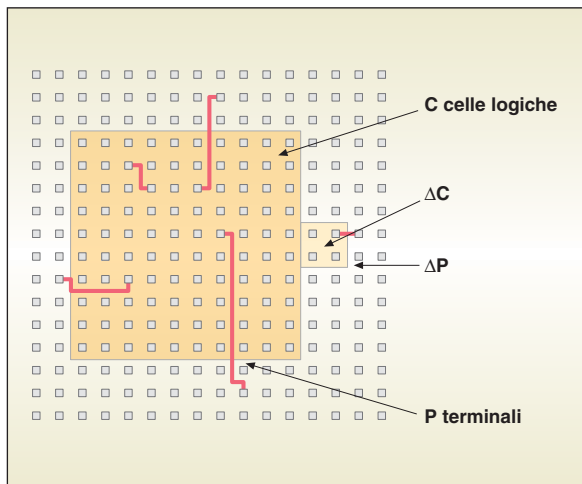


Fig. 1 – Introduzione di una perturbazione nella geometria della regione mediante l'aggiunta di un piccolo gruppo di porte logiche

stesso livello di connettività di quello dei gate del blocco iniziale.

Se C è il numero di celle logiche che compongono il blocco e P il numero di terminali che lo collegano al resto del sistema, l'introduzione della perturbazione comporta un'aggiunta di ΔC gate a cui sarà associata una variazione di ΔP terminali. Poiché dai dati iniziali si conosce che in media per ogni gate sono associati P/C terminali, un'estrapolazione lineare ci porta a concludere che:

$$\Delta P = \left(\frac{P}{C} \right) \Delta C$$

Se ΔP e ΔC sono di entità trascurabile rispetto a P e C la precedente equazione alle differenze origina la seguente equazione differenziale del primo ordine:

$$\frac{dP}{P} = \frac{dC}{C}$$

Ricordando che l'integrale di $1/P$ è il logaritmo di P , la soluzione vale:

$$P = k C$$

dove k è una costante d'integrazione che dipende dal tipo di chip e rappresenta il numero medio di terminali necessari

per ogni gate (tale interpretazione deriva dal fatto che quando $C=1$ si ha $k=P$). Quest'analisi fornisce una stima pessimistica in quanto non prevede il processo di ottimizzazione del piazzamento dei gate. Questo processo, sfruttando algoritmi avanzati come ad esempio il simulated annealing, modifica la posizione delle celle logiche all'interno del blocco in maniera da favorire le connessioni tra gate a breve distanza su quelle che avvengono tra gate distanti o appartenenti a blocchi diversi. Quindi il fatto che la perturbazione del blocco logico in esame sia avvenuta con l'aggiunta di un piccolo blocco attiguo porta a concludere che la precedente stima non sia accurata. Infatti, dopo il processo di ottimizzazione compiuto dagli algoritmi di piazzamento, sarà molto probabile che le connessioni che attraversano i confini della regione incrementale rientrino nei confini del blocco originale e che queste non contribuiscano ad incrementare il numero totale dei terminali del blocco perturbato. Questo livello di ottimizzazione viene rappresentato da un fattore correttivo indicato dal parametro $0 < p_i < 1$. L'equazione alle differenze si modifica nella seguente:

$$\Delta P = p_i \left(\frac{P}{C} \right) \Delta C$$

Approssimando questa equazione alle differenze alla corrispondente equazione differenziale e risolvendo con le medesime condizioni al contorno si ottiene:

$$P = k C^{p_i}$$

Tale equazione definisce in maniera matematica le conclusioni della regola di Rent ed è stata formulata nel 1971 dai due ricercatori dell'IBM Landman e Russo. La costante k è denominata coefficiente di Rent, l'esponente p_i esponente interno di Rent. Questo esponente rappresenta il livello di ottimizzazione del piazzamento raggiunto in una regione omogenea caratterizzata da un numero medio di k terminali per cella logica. Se $p_i = 1$, ovvero è pari al suo limite, non c'è ottimizzazione del piazzamento. In queste condizioni il blocco logico è interpretato come un insieme casuale di celle logiche e, quindi, di difficile interconnessione. Il limite inferiore di p_i è definito dalla topologia delle interconnessioni in quanto non è possibile garantire che tutte le celle da interconnettere siano piazzate in maniera adiacente. Tale valore, chiamato esponente intrinseco di Rent, è indicato con p^* (risulta $p^* < p_i < 1$).

Accanto all'esponente interno viene definito un esponente esterno di Rent, indicato con p_e . Se il numero di gate che compone la regione considerata tende a raggiungere il numero totale dei gate del circuito integrato, ovvero se il confine della regione tende ad essere quello del chip, il numero di terminali risulta limitato dalle porte di I/O dislocate in periferia. Da questo si evince un fenomeno di distorsione che tende a diminuire il numero dei terminali stimabili. Quando l'esponente è pari al valore 0,5 viene stabilito empiricamente che il numero di I/O di un circuito integrato è approssimativamente legato alla radice quadrata del numero delle sue celle logiche. È stato calcolato che per un proces-

sore Intel 8086 l'esponente esterno di Rent assume il valore 0,36; per un processore Motorola 68000 (68000 transistor e 64 terminali) vale 0,27; per un PowerPC IBM/Motorola (7 milioni di transistor e 482 pin) vale 0,32. La precedente formula diviene:

$$n_{I/O} = k G_T^{p_e}$$

dove $n_{I/O}$ è il numero di pin del chip e GT il numero totale dei gate.

Le informazioni desunte dall'esponente esterno di Rent sono difficilmente interpretabili se applicate al piazzamento delle celle logiche nel chip o alla qualità del routing, ovvero della procedura di interconnessione delle celle.

Generalmente si fa riferimento all'esponente interno che, nel seguito, verrà chiamato semplicemente esponente di Rent (k e p sono spesso chiamati parametri di Rent che, per le architetture di computer IBM analizzate inizialmente da Rent, valevano rispettivamente 2,5 e 0,6). Data la sua definizione esponenziale, l'esponente di Rent ha variazioni consistenti sulle grandezze da esso derivate come la media della lunghezza delle interconnessioni. È quindi opportuno averne una stima accurata. Un primo metodo è quello di utilizzare il valore desunto dai circuiti già realizzati.

La letteratura in materia è vasta e contraddittoria. Quasi tutte le fonti attribuiscono un campo di valori compreso tra $0,5 < p < 0,7$ (i valori calcolati da Landman e Russo erano $0,47 < p < 0,75$). Generalmente si riconosce che l'esponente di Rent sia basso per strutture regolari, come avviene per le memorie, i Gate Array e gli FPGA, e alto per strutture ad alta complessità, come avviene per gli ASIC e i computer. Nel primo caso, quando si analizzano architetture dotate di blocchi progettati per minimizzare il numero di pin, si attribuisce un valore di 0,5. Nel secondo caso, quando la complessità dell'architettura è paragonabile a logica sparsa si attribuiscono valori compresi tra 0,6 e 0,7. Da questi

TABELLA 1 - CARATTERIZZAZIONE DEI PARAMETRI DI RENT DOVUTA A BAKOGLU

Tipo di sistema	Esponente p	Coefficiente k
Memorie statiche	0,12	6
Microprocessori	0,45	0,82
Gate Array	0,50	1,9
FPGA	0,57 ÷ 0,62	*
Computer	0,63	1,4
ASIC	0,69	0,8
Board	0,25	82

(* dipende dall'architettura delle celle logiche)

dati si desume che l'esponente di Rent fornisce una misura della complessità delle interconnessioni di un dato circuito. Più tale valore è alto più grande sarà l'effort da spendere per realizzare il routing.

Per fronteggiare l'ambiguità di tali dati occorrerebbe misurare l'esponente di Rent direttamente dalla netlist finale. Ma, per ricavare tale file, è necessario un tempo di elaborazione elevato. Per cui è prassi comune stimarne il valore mediante un campionamento delle proprietà della netlist prima del piazzamento e del routing.

Partizionando opportunamente la logica si rientra in una classe di dati storici e di benchmark che, ad un dato numero di gate e per una data architettura, consente di esaminare grafici in scala logaritmica con associato il valore di tale esponente.

Applicazione della legge di Rent negli FPGA

Nella maggioranza dei casi la regola di Rent è stata utilizzata per fornire una stima a priori della lunghezza delle interconnessioni di un circuito integrato. Per una logica programmabile tali informazioni sono state utilizzate dai tool di timing-driven placement, da quelli di routing o da quelli che stimano la potenza dissipata. Anche se negli ultimi anni ci sono state evoluzioni nei processi di stima del routing, la regola di Rent continua ad essere applicata nella stima dell'area del routing. In tale ambito i

risultati applicati a diverse architetture di FPGA, o al confronto tra architetture di FPGA e di ASIC o structured ASIC, sono sorprendenti.

Utilizzando la regola di Rent, la stima della lunghezza delle interconnessioni si ottiene integrando la relazione di Rent sul numero totale di celle del circuito integrato.

$$W_{len} = \frac{1}{2} k \left[\frac{C^{(p+\frac{1}{2})}}{2p-1} + C^p \right]$$

W_{len} è la lunghezza totale stimata delle interconnessioni.

Da questa grandezza è possibile ricavare l'area occupata dal routing (Wire Area):

$$Wire Area = \frac{k_r W_{len} C_p}{\# layers}$$

dove $\# layers$ è il numero dei livelli di metallizzazione disponibili per il routing; k_r è un fattore di utilizzazione del routing (tipicamente vale 2); C_p è l'area media di una connessione che attraversa una singola cella desunta dalla seguente relazione:

$$C_p = \frac{area_{cell}^{1,5}}{\left[\frac{connessioni verticali \times connessioni orizzontali}{\# layers^2} \right]}$$

Esisterà un punto nel quale il routing tra celle eccede l'area occupata dalle stesse.

Break Area = Max (Gate Area, Wire Area)
Questa relazione suggerisce che, per

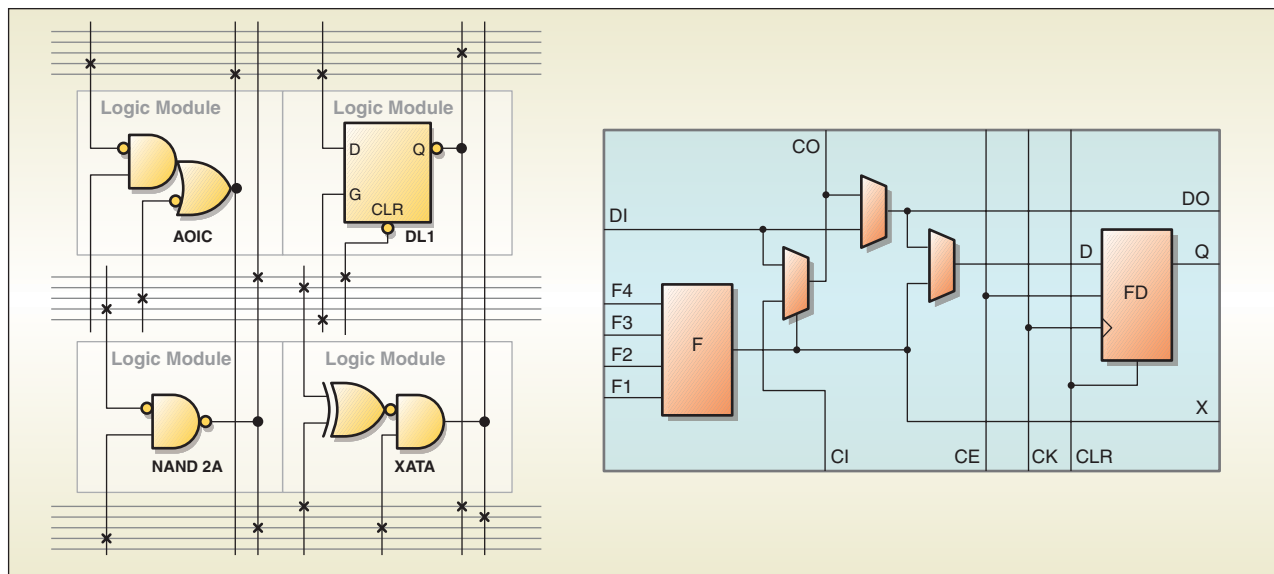


Fig. 2 – Esempio di cella logica fine-grained (a sinistra) e coarse-grained (a destra)

ogni dato numero di layer di routing, esiste una dimensione di progetto che eccede lo spazio riservato al routing nei confronti di quello disponibile. Ogni progetto che supera questa dimensione si dirà dominato dall'area del routing. In caso contrario, quando questo limite non è superato, si dirà dominato dall'area delle celle.

Questo modello fornisce un meccanismo per prevedere gli effetti dell'area tra un'architettura FPGA fine-grained e una coarse-grained. Un'architettura FPGA fine-grained, simile a quella degli ASIC, è costituita da celle logiche di tipo elementare, generalmente costituite da pochi transistor, circondate da numerose risorse di routing. Un'architettura FPGA coarse-grained, simile a quella degli structured ASIC, è costituita da celle logiche formate da uno o più flip-flop e una o più LUT circondate da risorse di routing di tipo gerarchico (LUT è acronimo di Look-Up Table che essenzialmente è una funzione combinatoria generalizzata di 4 o 5 variabili). In figura 2 si mostra un esempio di cella logica fine-grained (a sinistra) e coarse-grained (a destra).

È possibile approssimare una cella fine-grained a una NAND a tre ingressi. Per ogni cella, di area $7,5 \text{ micron}^2$, si assegna un valore medio di 2,5 pin di ingresso e 1 pin di uscita. Analogamente è possibile ritenere che una cella coarse-grai-

ned sia basata su una cella di tipo LUT alla quale vengono mediamente attribuiti un'area equivalente di 15 gate, pari a 180 micron^2 , 7,5 pin di ingresso e 1,5 pin di uscita.

Per entrambi si adotti il medesimo esponente di Rent (pari a 0,67).

Allo scopo di rendere il confronto plausibile si supponga che le architetture siano dotate di 4 piani di routing, cioè 4 layer, e che per compensare l'effetto del routing locale all'interno della cella di tipo coarse-grained vengano attribuiti due layer in più all'architettura fine-grained. Quindi ne risulta un confronto tra un'architettura coarse-grained a 4 livelli di metallizzazione e un'architettura fine-grained a 6 layer.

Superato il limite in cui l'area del routing è dominante, un dispositivo coarse-grained, contrariamente alle attese, è più denso di un fine-grained. Ciò è dovuto al maggior numero di pin per cella. Infatti, applicando la regola di Rent, si desume un valore del pinout di 21,5 pin,

molto superiore al valor medio dei pin della cella coarse-grained pari a 9. La natura di questo risultato garantisce la possibilità di avere un maggior numero di celle logiche a parità di area. L'ago della bilancia propende a favore degli FPGA del tipo Virtex o Stratix, e ancor più degli ASIC strutturati in grado di offrire consumi ridotti e die più piccoli, a discapito degli altri tipi di FPGA e degli ASIC.

Infine è da sottolineare che una cella logica composta da 15 gate di tipo fine-grained a 9 terminali ha un esponente di Rent pari a 0,35, ben al di sotto della logica sparsa caratterizzata da un esponente pari a 0,67, evidenziando un maggiore grado di regolarità. Nei casi in cui l'area del routing di un progetto eccede l'area delle celle logiche si è affermato che le architetture FPGA coarse-grained e structured ASIC sono più efficienti.

Tuttavia, è possibile affermare che tra le due, a parità di area di silicio, le seconde sono ancora più dense in virtù del fatto che non necessitano di interconnessioni programmabili. Uno switch programmabile, essenzialmente un latch e un pass transistor, è equivalente in termini di area a una porta logica di tipo nand e questa, integrata su tutte le risorse di routing disponibili, determina una penalità di area riducendo la densità del circuito integrato di tipo FPGA di circa un fattore 50.

✎