Définition d'une Métrique d'Insertion de Buffers

Xavier Michel, Alexandre Verle, Nadine Azemard, Philippe Maurine, Daniel Auvergne

Définition d'une métrique d'insertion de buffers

X. Michel, A. Verle, N. Azémard, P. Maurine, D. Auvergne

LIRMM, UMR CNRS/Université de Montpellier II, (C5506),
161 rue Ada, 34392 Montpellier, France
azemard@lirmm.fr

RESUME

La conception de circuits de haute performance exige un bon compromis entre la vitesse, la puissance et la surface. Basé sur une modélisation analytique du retard, ce travail présente une méthode pour définir des métriques permettant de caractériser la criticabilité des nœuds d'un circuit. Le but de ce travail est de définir des indicateurs afin de choisir entre les différentes solutions d'optimisation. La détermination de ces indicateurs est validée par comparaison avec les limites de charges obtenues à partir de simulations SPICE. L'application sur différents circuits ISCAS prouve que, sans énumération, une première amélioration du retard d'un chemin peut être obtenue avec un coût réduit en surface et en puissance.

1. INTRODUCTION

La conception de circuits de haute performance exige de définir un compromis entre la vitesse, la puissance et la surface. Ceci peut être obtenu en dimensionnant les transistors [1], ou en utilisant des techniques plus générales d'insertion de buffers [2] et de transformation logique [3]. La plupart des systèmes de synthèse emploient des arbres de buffers pour accélérer les chemins critiques. Bien que ces techniques soient efficaces pour accélérer les chemins combinatoires, elles influencent la dissipation résultante de puissance et la surface. Le dimensionnement des portes est coûteux en puissance et, en raison des effets des capacités de charge, peut ralentir les chemins adjacents. Ceci rend le problème complexe et nécessite un grand nombre d'itérations. L'insertion de buffer préserve les chemins, mais est efficace seulement pour des nœuds fortement chargés. L'insertion d'un seul inverseur peut être une solution intermédiaire, mais la conservation du signal logique implique une transformation. Pour choisir entre ces différentes alternatives, il est nécessaire d'évaluer et de comparer les performances de chaque décomposition ou alternative d'assignement, ce qui implique une interprétation temporelle correcte des performances de la librairie dans la technologie utilisée.

Sans indicateur robuste, le choix entre ces différentes techniques, pour toutes les portes d'une bibliothèque, est complexe. Les solutions itératives explosent en temps CPU. Un choix raisonnable de technique d'accélération doit être basé sur la détermination des noeuds critiques, et la caractérisation de la sensibilité des portes au dimensionnement et à la bufferisation.

Différentes techniques sont utilisées pour contrôler les fanouts importants dans les circuits. Elles utilisent soit le dimensionnement des transistors, soit l'insertion de buffers, soit les deux combinés. Le problème du dimensionnement global des transistors est souvent résolu en utilisant des techniques de programmation non linéaires [1]. Le dimensionnement discret des transistors ou le dimensionnement d'une bibliothèque spécifique, dans laquelle seulement un nombre limité de tailles est disponible pour chaque porte, est un problème d'optimisation NP complet [4], résolu habituellement par des heuristiques [5]. La plupart de ces techniques emploient un modèle simplifié du délai d'Elmore. Elles sont basées sur des itérations successives et appliquées à toutes les portes du chemin étudié. Sutherland [6] minimise le retard sur un chemin en imposant le même effort à chaque étage de ce chemin. Cette approche minimise le retard de commutation pour une succession idéale de portes (c'est à dire sans capacité parasite), mais ne permet pas d'avoir une solution optimale en délai et en surface pour des structures réelles. Cependant aucun indicateur n'existe pour identifier les noeuds critiques ou pour choisir la technique la plus efficace pour accélérer la porte correspondante, dite critique.

Afin d'aider les concepteurs à identifier les portes critiques et à choisir parmi les différentes optimisations, nous définissons dans cet article une métrique pour l'insertion de buffers, utilisable

comme un indicateur efficace pour caractériser les portes logiques en termes de sensibilité aux techniques de dimensionnement et de bufferisation.

Dans cet article, nous présentons d'abord le modèle de délai [7], et nous discutons des conditions d'insertion de buffers. Dans la section 3, nous définissons la charge limite à considérer pour une porte pour l'optimisation de son fanout. La validation de ces limites est présentée dans la section 4 et la conclusion dans la section 5.

2. MODELE DE DELAI

2.1 Modélisation au niveau transistor

Le modèle de délai [7] au premier ordre que nous utilisons est la généralisation de la réponse indicielle du modèle de Mead [8] :

$$t_{HLstep} = \frac{C_L \cdot \Delta V}{I_N} = \tau \cdot \frac{C_L}{C_N}$$
$$t_{LHstep} = \frac{C_L \cdot \Delta V}{I_P} = \tau \cdot R \cdot \frac{C_L}{C_P} \quad (1)$$

où τ est une unité de temps qui caractérise le process et peut être calculée ou directement déterminée par simulation de ce process. C_L, C_N et C_P représentent, respectivement, la charge de sortie et la capacité des transistors N et P de la porte considérée. τ est défini pour le front descendant. R représente, pour des capacités de charge et de drive identiques, le courant dans les transistors N et P.

Suivant [6], l'extension aux portes est obtenue en ramenant chaque porte à un inverseur équivalent. Nous considérons le pire cas. Les possibilités en courant des transistors N (P) en parallèle sont équivalentes au courant maximum d'un inverseur avec des transistors de tailles identiques. Le réseau série de transistors N (P) est modélisé comme un générateur de courant contrôlé par une tension d'entrée avec des possibilités en courant réduites par un facteur DW. Ce facteur de réduction (DW) est défini comme le rapport du courant disponible dans un inverseur à celui d'un réseau de transistors en série. Ceci résulte en une expression plus générale de l'expression (1) pour une porte :

$$t_{HLstep} = \tau \cdot S_{HL} \cdot \frac{C_L}{C_{IN}}$$
$$t_{LHstep} = \tau \cdot S_{LH} \cdot \frac{C_L}{C_{IN}} \quad (2)$$

où $C_{IN} = C_N(1+k)$ est la capacité d'entrée de la porte, et k représente le rapport de configuration entre transistors N et P.
Pour simplifier, le facteur S inclut toute la différence de possibilités en courant entre les transistors équivalents du plan N et du plan P qui sont dépendants du rapport de configuration et de la saturation en vitesse :

$$S_{HL} = \frac{(1+k)}{2} \cdot DW_{HL}$$
$$S_{LH} = \frac{(1+k) \cdot R}{2k} \cdot DW_{LH} \quad (2a)$$

où DW représente, pour chaque front, le rapport de courant admissible dans un inverseur et une porte de taille identique :

$$DW_{HL,LH} = \frac{I_{N,P}(Inv)}{I_{N,P}(Gate)} \cdot \frac{W_{N,P}(Inv)}{W_{N,P}(Gate)} \quad (2b)$$

Ces facteurs de réduction représentent de façon explicite l'effort logique présenté dans [6].
Un exemple de la valeur de ces coefficients est donné dans le tableau 1. Ils ont été déterminés pour une technologie de 0.25µm pour des portes avec un rapport de configuration k=1.

1.2 Calcul du délai

Un calcul réel du délai doit prendre en compte les transitions d'entrée. Comme développé dans [9], nous introduisons l'effet de couplage entrée-sortie et les effets de rampe d'entrée sur le modèle par :

$$t_{HL}(i) = \frac{v_{TN}}{2} \tau_{INLH}(i-1) + (1 + \frac{2C_M}{C_M + C_L}) t_{HLstep}(i)$$
$$t_{LH}(i) = \frac{v_{TP}}{2} \tau_{INHL}(i-1) + (1 + \frac{2C_M}{C_M + C_L}) t_{LHstep}(i) \quad (3)$$

où $\tau_{INHL,LH}$ est la durée du signal d'entrée égal au double du temps de réponse de la porte contrôlante. C_M est la capacité de couplage entre les nœuds d'entrée et de sortie, pouvant être évaluée comme la moitié de la capacité d'entrée du transistor P(N) pour le front montant (descendant) respectivement, ou directement calibrée à partir de simulation SPICE.

Dans (3), nous avons considéré que le temps de transition en entrée est assez court pour permettre de considérer que la commutation de la porte en sortie se produit toujours avec un courant maximum et constant. Ceci justifie l'utilisation des temps de réponse pour évaluer le retard.

3. OPTIMISATION DE FANOUT

3.1 Alternatives d'optimisation de Fanout

Nous adressons ici le problème de définir un indicateur, permettant de choisir entre diverses alternatives d'optimisation. Considérons un chemin critique, défini comme un chemin ne respectant pas la contrainte de délai, nous proposons d'identifier les noeuds critiques et de choisir alors la meilleure technique d'optimisation. Une autre alternative, avant toute distribution de budget sur le chemin, est que l'identification des noeuds critiques, permettra une optimisation locale pour améliorer l'algorithme de distribution de la contrainte. À ce niveau, trois actions bien identifiées peuvent être considérées.

La première solution est le dimensionnement des transistors. Augmenter la taille des transistors de la porte contrôlant le noeud critique, diminue la charge de celle-ci de par le rapport (CL/CIN) de cette porte (1). Si cette solution améliore le retard de cette porte, elle affecte la vitesse de la porte précédente et, en raison de l'effet de rampe d'entrée, réduit le gain initial.

Les deuxième et troisième solutions consistent à insérer un buffer entre la porte et sa charge de sortie. L'utilisation d'un seul inverseur impose une reconfiguration logique pour conserver la parité du signal. Ceci peut être intéressant si une reconfiguration plus rapide peut remplacer la porte fortement chargée. L'insertion de deux inverseurs (buffer) améliore la vitesse seulement pour des charges importantes. L intérêt significatif en insérant un ou deux inverseurs est d'accélérer la commutation de la porte sans modifier sa capacité d'entrée. Une solution possible pour choisir la meilleure alternative d'accélération, sans aucune itération, est de caractériser chaque porte et sa charge en termes de criticabilité.

3.2 Définition d'une métrique

Nous voulons développer une méthode permettant d'évaluer le niveau de charge des différents noeuds. Pour cela, nous utilisons le facteur de Fanout F0=CLi/CINi pour évaluer le niveau de charge d'une porte (i). CLi représente sa charge de sortie et CINi sa capacité d'entrée. Nous caractérisons chaque type de porte par un facteur de fanout spécifique, F0lim, pouvant être utilisé aussi bien comme indicateur de la sensibilité de la porte au dimensionnement ou à l'insertion de buffers, que comme seuil pour choisir entre les alternatives d'optimisation. Nous considérons des portes NAND et NOR simples et des inverseurs avec les transistors N et P identiques (k=1). D'autres valeurs du rapport de configuration k peuvent être considérées ce qui augmente les facteurs de réduction comme indiqué dans le tab. 1.

	Facteurs de réduction S	
Portes, 0.25µm,	S_{HL}	S_{LH}
Inverseur k = 1	1	2.3
Inverseur k = 2	1.5	1.73
Inverseur k = 3	2	1.53
Nand2 k = 1	1.55	2.3
Nand3 k = 1	2.05	2.3
Nor2 k = 1	1	4.3
Nor3 k = 1	1	6.3

Tableau 1 : Exemple de valeurs des facteurs de réduction pour des portes simples en CMOS 0.25µm.

Dans la Fig.1, sont représentées les trois configurations à considérer pour choisir entre dimensionnement et bufferisation : (a) une porte avec une charge de sortie C_L , (b) et (c) la même porte après insertion d'un ou de deux inverseurs.

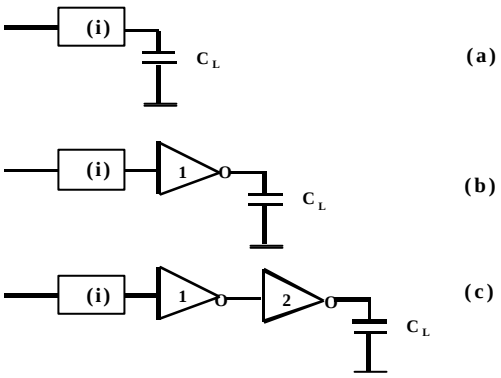

Figure 1 : Structures à considérer pour déterminer les limites d'insertion de buffers.

Il apparaît clairement que, pour chaque porte considérée, l'implantation à minimum de délai dépend de la valeur de la charge de sortie et du dimensionnement des inverseurs. L'insertion d'un inverseur correctement dimensionné, distribue la charge initiale et accélère la porte au détriment du temps de propagation. La valeur limite de sortie à partir de laquelle les solutions (b) ou (c) sont plus rapides que l'implantation initiale, est un bon indicateur de la criticabilité de la porte étudiée. Ceci est illustré dans la Fig.2, qui représente l'évolution

du délai en fonction de la charge d'une porte Nor 2 entrées. L'insertion d'un (1) ou deux inverseurs (2) a une influence significative sur la sensibilité à la charge de la structure. L'intersection entre les courbes définit les fanout limites permettant de choisir entre dimensionnement et insertion de buffer.

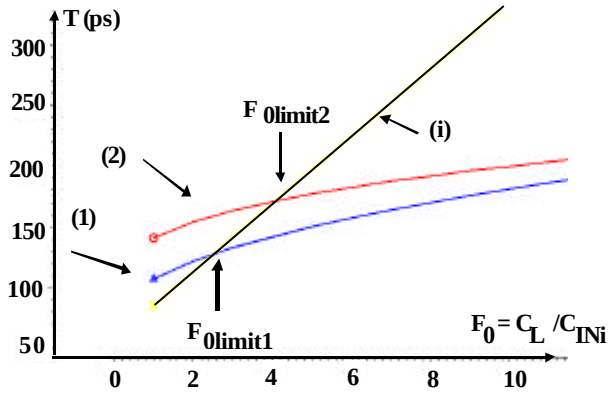

Figure 2 : Sensibilité du délai pour une porte Nor 2 entrées (avec une capacité d'entrée de 10.4fF) pour les 3 configurations de la figure 1.

Maintenant, calculons cet indicateur. En utilisant (3), pour évaluer le retard de l'entrée vers la sortie des structures de la Fig 1, il est simplement nécessaire de déterminer la charge de sortie C_L satisfaisant les inégalités suivantes :

$$\begin{aligned} t_{HL,LH}(a) &\geq t_{HL,LH}(b) \\ t_{HL,LH}(a) &\geq t_{HL,LH}(c) \end{aligned} \qquad (4)$$

Le développement de (4) va permettre de dimensionner correctement les inverseurs des configurations b et c. Pour cela, nous imposons les contraintes suivantes :

- le dimensionnement de la porte (i) doit être inchangé,
- les retards de propagation (HL,LH) de la nouvelle structure doivent être identiques entre eux et minimisés.
- la solution dimensionnée doit être l'alternative à surface la plus faible. Cette surface est évaluée en tant que somme des largeurs des transistors, ce qui est équivalent, pour simplifier, à la somme des capacités d'entrée correspondante.

De plus la solution de dimensionnement dépend de la porte initiale (i) considérée.

A – Insertion d'un inverseur.

Si la porte initiale est un inverseur, il est facile de montrer à partir de (1) et (3) que l'égalité des temps de montée et de descente dans un réseau de deux inverseurs est obtenue quand ils ont des rapports de configuration identiques (ki = k1 = k dans Fig.1b). Nous imposons, aussi, la même condition pour les portes NAND et NOR. Alors, la capacité d'entrée de l'inverseur inséré, est obtenue à partir de :

$$C_{IN1} = \sqrt{\frac{SW_{LH1} - SW_{HL1}}{SW_{LHi} - SW_{HLi}}} \sqrt{C_L \cdot C_{INi}} \qquad (5)$$

Ici, pour minimiser la surface totale, le rapport de configuration de chaque élément est égal à 1, aussi longtemps que la racine carrée de (5) reste positive.

B – Insertion de 2 inverseurs (Bufferisation).

Nous voulons imposer la même contrainte de symétrie du délai sur la structure, tout en minimisant la surface totale. Avec un nombre impair d'étages (configuration (c)), l'unique solution est d'imposer, comme dans la partie précédente, l'égalité des retards sur deux étages et la symétrie des fronts sur l'étage restant. Dans la configuration (c), la solution efficace en surface consiste à imposer l'égalité des délais entre les deux premières étages, (i) et (1), et d'équilibrer les fronts de l'étage (2) en choisissant k2 = R. Comme donné dans (1), ceci est la condition, sur le rapport de configuration de l'inverseur, pour équilibrer les fronts de montée et de descente.

Cette fois, deux paramètres doivent être déterminés, C_{IN1} et C_{IN2}. La première équation est obtenue en appliquant (5) aux étages (i) et (1), ce qui donne :

$$C_{IN1} = \sqrt{\frac{SW_{LH1} - SW_{HL1}}{SW_{LHi} - SW_{HLi}}} \sqrt{C_{IN2} \cdot C_{INi}} = A \cdot \sqrt{C_{IN2} \cdot C_{INi}} \qquad (6)$$

L'expression de délai résultante dépend seulement de C_{IN2}. La dérivée de ce délai par rapport à C_{IN2} est égale à zéro pour :

$$C_{IN2} = \left[\frac{(1+R)}{\frac{1}{2}\left(SW_{HLi} \cdot A + \frac{SW_{HL1}}{A} \right)} \right]^{2/3} \cdot C_{INi}^{1/3} \cdot C_L^{2/3} \qquad (7)$$

$$C_{IN1} = A \cdot \left[\frac{(1+R)}{\frac{1}{2}\left(SW_{HLi} \cdot A + \frac{SW_{HL1}}{A} \right)} \right]^{1/3} \cdot C_{INi}^{2/3} \cdot C_L^{1/3} \qquad (8)$$

Nous avons mis en application ces valeurs de la capacité d'entrée de l'inverseur dans les équations du retard des configurations (b) et (c) de la Fig.1, pour obtenir la charge de sortie satisfaisant (4). Le facteur de fanout résultant caractérise le niveau de charge de la porte au delà duquel n'importe quelle insertion d'inverseur (buffer) accélère plus efficacement la porte que n'importe quel dimensionnement de ses transistors.

4. RESULTATS EXPERIMENTAUX

Dans le tableau 2, nous comparons les valeurs simulées (HSPICE 0.25µm, level 49) et calculées du facteur de charge limite pour différentes portes. Nous avons défini par p = Cpar/CIN la capacité parasite de chaque porte. et avons considéré deux conditions de charge parasites (p = 0 ou 1). Nous obtenons une bonne concordance entre les valeurs calculées et simulées.

Nous devons noter ici que le fanout limite, obtenu pour l'insertion de buffer sur un noeud contrôlé par un inverseur ou une porte NAND, correspond à une charge assez grande. Mais pour une porte NOR, la limite est plus faible ; ceci indique que cette catégorie de porte ne peut accepter qu'une charge égale à une à deux fois sa capacité d'entrée. De tels noeuds seront des noeuds privilégiés pour l'insertion d'inverseur ou de buffer. En considérant, par exemple, la limite d'une porte NOR 3 pour insérer un inverseur, nous pouvons facilement conclure que n'importe quelle synthèse logique basée sur un OR à 3 entrées sera plus rapide qu'en utilisant des portes NOR.

L'utilisation de ce facteur de charge limite donne directement sur un chemin, sans itération, la liste des portes susceptibles d'être optimisées, ce qui permet de définir le protocole d'optimisation suivant:

- Si $F_0 < F_{0limit}$: dimensionnement des transistors de la porte considérée, ceci sera utilisé pour la distribution de contraintes sur un chemin, (voir [5] par exemple).
- Si $F_0 > F_{0limit}$: insertion de 1 ou 2 inverseurs. un chemin. Ainsi, la distribution d'une contrainte de délai sur le chemin étudié donnera une implantation à surface plus faible.

porte		F_{0lim} 1 inverseur p=0	1 inverseur p=1	2 inverseurs p=0	2 inverseurs p=1
INV k = 1	Simul.	3.44	4.46	5.7	7.5
	Calcul.	3.36	4.44	5.2	7.2
INV k = 2	Simul.	6.05	7.97	9	12.5
	Calcul.	5.82	7.6	8	11
INV k = 3	Simul.	5.1	7	7.3	10.3
	Calcul.	4.9	6.8	7.3	9.9
NAND2	Simul.	4.1	5.2	6.1	8.6
	Calcul.	4.3	5.6	5.6	7.8
NAND3	Simul.	4.6	5.7	6.6	9.4
	Calcul.	4.5	6	6.1	8.4
NOR2	Simul.	1.7	2.35	2.9	4.1
	Calcul.	1.7	2.4	3	4.1
NOR3	Simul.	1.1	1.46	1.8	2.5
	Calcul.	1.14	1.48	2.4	3.2

Tableau 2 : Comparaison des valeurs simulées et calculées du facteur de charge limite.

Ce protocole a été appliqué sur différents circuits ISCAS. Certains résultats sont représentés dans le tableau 3 où nous comparons la surface (en tant que somme des largeurs des transistors) des différents circuits, avant et après bufferisation. Pour l'implantation initiale, tous les transistors sont à la largeur minimale. La contrainte de délai a été satisfaite en utilisant le logiciel AMPS de chez Synopsys. Comme nous pouvons le voir dans ce tableau, une insertion contrôlé de buffers entraîne un gain en surface intéressant.

Cette application sur différents circuits ISCAS prouve que, sans énumération, une première amélioration du retard d'un chemin peut être obtenue avec un coût réduit en surface et en puissance.

		Avant bufferisation			Après bufferisation
	Contrainte de délai T(ns)	Délai initial (ns)	Surface initiale ΣW (µm)	Surface à T contraint (µm)	Surface à T contraint (µm)
C18 3 portes	1.3	3.5	4.2	50.4	7.9
FAPD 8 portes	0.6	1.5	8.4	49	34.5
FPD 13 portes	0.8	1.8	13	70.4	41.1

Tableau 3 : Gain en surface après bufferisation sur différents circuits ISCAS.

5. CONCLUSION

Dans cet article, nous avons proposé une nouvelle définition de métrique pour qualifier le niveau de charge critique des nœuds de chemins combinatoires. Des indicateurs pour l'insertion de buffer ont été définis en fonction des paramètres physiques de conception. Pour les différentes portes d'une bibliothèque, on a obtenu les limites de charge de sortie, ce qui permet de classer les portes en terme de sensibilité de charge et bien plus de choisir entre les alternatives de dimensionnement ou d'insertion de buffer, sans itération. Résoudre les nœuds fortement chargés avant toute distribution de contrainte en délai, apparaît comme une solution très intéressante pour satisfaire sur un circuit une contrainte en délai avec le plus petit coût en surface/puissance.

6. REFERENCES

[1] J. M. Shyu, A. Sangiovanni-Vincentelli, J. Fishburn, A. Dunlop, "Optimization-based transistor sizing" IEEE J. Solid State Circuits, vol.23, n°2, pp.400-409, 1988.

[2] S.R. Vemuru, A.R. Thorbjornsen, A.A. Tuszynski, " CMOS tapered buffer", IEEE J. Solid State Circuits, vol.26, n°9, pp.1265-1269,1991.

[3] P.G. Paulin, F. J. Poirot, "Logic decomposition algorithm for the timing optimization of multilevel logic", Proc. ICCD 89, pp.329-333.

[4] P. K. Chan, "Algorithms for library-specific sizing of combinational logic" in Proc. DAC, 1990, pp.353-356.

[5] J. Fishburn, A. Dunlop, "TILOS: a posynomial programming approach to transistor sizing" in Proc. Design Automation Conf. 1985,pp.326-328.

[6] I. Sutherland, B. Sproull, D. Harris, "Logical Effort: Designing Fast CMOS Circuits", Morgan Kaufmann Publishers, INC., San Francisco, California, 1999.

[7] D.Auvergne, J.M. Daga, M. Rezzoug "Signal transition time effect on CMOS delay evaluation" IEEE trans. on Circuits and Systems: Fundamental theory and applications, vol.47, n°9,pp.1362-1369, sept.2000.

[8] C. Mead, M. Rem, "Minimum propagation delays in VLSI", ", IEEE J. Solid State Circuits, vol.SC17, n°4, pp.773-775, 1982.

[9] K.O. Jeppson, "Modeling the influence of the transistor gain ratio and the input-to-output coupling capacitance on the CMOS inverter delay", IEEE J. Solid State Circuits, vol.29, pp.646-654, 1994.