Modélisation, Estimation et Optimisation de la consommation des interconnexions dans les SOC

Antoine Courtay, Johann Laurent, Nathalie Julien, Olivier Sentieys

Modélisation, Estimation et Optimisation de la consommation des interconnexions dans les SOC

Antoine Courtay[1,2], Johann Laurent[1], Nathalie Julien[1], Olivier Sentieys[2]
[1]Lab-STICC - Université de Bretagne Sud - rue saint Maudé, 56100 Lorient, France
prenom.nom@univ-ubs.fr
[2]IRISA - Université de Rennes 1 - 6, rue de Kerampont, 22300 Lannion, France
nom@irisa.fr

1. INTRODUCTION

Les SOC d'aujourd'hui sont de plus en plus complexes et recquierent donc beaucoup de ressources de calcul ; ce qui implique un fort volume de données à stocker en mémoire ou à communiquer entre les unités de calcul. Pour transmettre ces données entre les unités (mémoires, processeurs ...) des bus ou des réseaux sur puce sont utilisés.
Dans les SOC actuels, la consommation due aux interconnexions peut représenter jusqu'à 50 % de la consommation totale de la puce [1]. De plus, la réduction des dimensions des transistors et des fils pose un problème de plus en plus important au niveau du temps de propagation et de l'énergie dus aux interconnexions [2].
Beaucoup de travaux, à différents niveaux d'abstraction autour de l'optimisation de la consommation des interconnexions ont été proposés. Malheureusement, la plupart des techniques d'optimisation proposées dans la littérature ne sont plus efficaces en terme de réduction de la consommation. Ceci est dû au fait que les longueurs d'interconnexion que l'on trouve actuellement dans les SOC sont plus courtes qu'avant. Il devient donc indispensable de prendre en compte la consommation des interconnexions lors de l'évaluation de la consommation d'une puce. Pour ce faire il est nécéssaire d'utiliser des outils d'estimation de la consommation des interconnexions lors des premières phases de l'élaboration d'un système. Il est également nécéssaire de proposer de nouvelles techniques d'optimisation de la consommation des interconnexions répondant aux spécificités des interconnexions actuelles.
La suite de cet article présentera dans un premier temps les paramètres à prendre en compte lors de la modélisation (délai et consommation) des interconnexions, ainsi que les modèles de consommation développés. La troisième partie présentera les résultats expérimentaux obtenus sur l'évaluation des techniques de réduction de la consommation à l'aide de l'outil. Basée sur les résultats expérimentaux cette partie proposera également les pistes à suivre pour les optimisations futures.

2. MODÉLISATION ET ESTIMATION DE LA CONSOMMATION

2.1 Modélisation de la consommation des bus

La consommation des interconnexions est liée à deux paramètres principaux qui sont la résistance et la capacité des lignes d'interconnexions.
Les grandeurs élémentaires que l'on trouve dans les $DesignKit$ des constructeurs permettent de calculer la résistance (R) ainsi que la capacité (C) du fil en fonction de ses dimensions. Afin de modéliser le plus précisément le fil, un modèle où R et C sont distribués sera utilisé afin d'avoir une meilleure précision au niveau du délai. Généralement, les lignes sont regroupées afin d'obtenir un bus de données permettant la transmission d'information entre blocs. Le fait d'utiliser un bus fait apparaître un autre phénomène capacitif qui est le couplage capacitif entre fil aussi appelé $crosstalk$. La capacité de $crosstalk$ dépend quant à elle de la surface en regard entre les fils et varie donc en fonction des dimensions technologiques (épaisseur et espacement). Le bruit dû au $crosstalk$ est relativement localisé ; en général un système soumis au $crosstalk$ est modélisé en négligeant les ordres supérieurs au premier. La figure 1 présente le modèle physique du bus que nous avons utilisé pour mener nos expérimentations. Un des effets du $crosstalk$ est de faire augmenter le délai de propagation sur le bus, il introduit un facteur de délai g comme décrit dans le tableau 1 où r est le ratio de la capacité de $crosstalk$ C_c par rapport à la capacité du fil C_s (capacité du fil par rapport au substrat).
Dans ce tableau, $\uparrow$ représente une transition montante, $\downarrow$ représente une transition descendante et $-$ signifie qu'il n'y a pas de transition sur le fil. Dans le meilleur cas lorsque les trois fils effectuent une transition dans le même sens, le délai sur le fil victime (le fil central) est le délai sans $crosstalk$ (i.e. $g = 1$) ; mais le cycle d'horloge doit être dimensionné en tenant compte exclusivement du délai pire cas (i.e. $g = 1 + 4.r$) afin d'assurer une transmission

Figure 1: Modèle complet $\pi 3$ pour 3 fils avec couplage *crosstalk*.

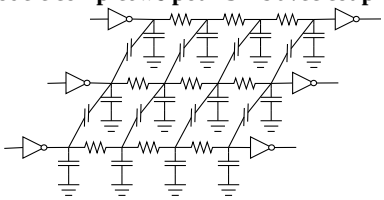

Table 1: Capacité parasite (C_L) et facteur de délai (g) du fil victime en fonction du type de transition.

C_L	Types de transition				g
C_s	$(\uparrow,\uparrow,\uparrow)$	$(\downarrow,\downarrow,\downarrow)$			1
$C_s + C_c$	$(-,\uparrow,\uparrow)$	$(-,\downarrow,\downarrow)$	$(\uparrow,\uparrow,-)$	$(\downarrow,\downarrow,-)$	1+r
$C_s + 2.C_c$	$(-,\uparrow,-)$ $(\uparrow,\uparrow,\downarrow)$	$(-,\downarrow,-)$ $(\uparrow,\downarrow,\downarrow)$	$(\downarrow,\uparrow,\uparrow)$	$(\downarrow,\downarrow,\uparrow)$	1+2.r
$C_s + 3.C_c$	$(-,\uparrow,\downarrow)$	$(-,\downarrow,\uparrow)$	$(\uparrow,\downarrow,-)$	$(\downarrow,\uparrow,-)$	1+3.r
$C_s + 4.C_c$	$(\uparrow,\downarrow,\uparrow)$	$(\downarrow,\uparrow,\downarrow)$			1+4.r

correcte des données.
Le $crosstalk$ représente également une source de bruit ; en effet, à cause de la capacité introduite entre les fils, une transition sur un des fils à pour effet de causer un pic de tension sur le fil voisin ce qui peut conduire à des erreurs à la réception si ce pic dépasse la tension de seuil des buffers.

2.2 Estimation de la consommation

Afin d'établir nos modèles de consommation du bus, nous avons besoin de connaître les caractéristiques physiques des interconnexions. Pour cela nous avons effectué les différentes expérimentations avec un simulateur SPICE (ELDO 5.7) pour plusieurs technologies ($130nm$, $90nm$ et $65nm$) afin d'obtenir des modèles précis au niveau physique en terme de consommation et de vitesse de transmission des données. Au terme des expérimentations, nous avons développé un outil qui fournit à l'utilisateur des résultats en termes de surface, de consommation (instantannée, moyenne et maximale), d'activité des données et de vitesse maximale de transmission. Ces résultats permettent alors à l'utilisateur de dimensionner son bus. L'outil va également permettre d'analyser efficacement les performances des différentes techniques de réduction de la consommation et du délai trouvées dans la littérature.

3. IMPACT DES TECHNIQUES D'OPTIMISATION ET PISTES À SUIVRE

On trouve dans la littérature beaucoup de techniques qui visent à réduire l'effet du $crosstalk$ ayant pour but l'accélération de la propagation des informations et également la réduction de la consommation des interconnexions. Ces techniques interviennent à plusieurs niveaux d'abstractions.
C'est au niveau architectural que l'on rencontre le plus grand nombre des techniques [3, 4, 5, 6]. Elles consistent toutes en un codage des données dont le principe est de transmettre l'information sur m bits avec ($m \geq n$) tel que l'activité des données codées soit inférieure à celle des données non codées. Le codage des données ayant pour but de supprimer les pires cas du tableau 1 (i.e. $g = 1 + 4.r$ et/ou $g = 1 + 3.r$).
Les résultats présentés dans [7] montrent que la classification des transitions (selon l'importance de la capacité du fil victime) diffère selon que l'on regarde la consommation ou le delai sur les interconnexions. D'un point de vue consommation, la classification des transitions (de la moins à la plus consommatrice) est d'abord commencée par les transitions montantes suivie des descendantes. Par conséquent, une première piste pour l'optimisation de la consommation est de coder les données tel que les transitions descendantes sur le bus soient celles qui présentent la plus faible capacité et donc la consommation énergétique la plus faible.
Dans un second temps, en analysant le profil d'activité des stimuli de type données (image, musique , parole), il a été remarqué que, appliquer les techniques d'optimisation sur les bits de poids faible permet d'obtenir de meilleurs résultats en terme de réduction de la consommation des interconnexions. Ceci étant dû au fait que les bits de poids faible présentent une activité importante comme démontré dans [8]. Par exemple, la technique du $PartialBusInvert$ permet d'obtenir de meilleurs résultats que le $BusInvert$.
Troisièmement, en analysant le pourcentage d'apparition des différentes classes de transition, il a été noté que les classes de transition éliminées par les techniques d'optimisation n'apparaissent pas très souvent. Par exemple, éliminé les deux dernières classes de transition comme celà est souvent proposé pour l'optimisation du delai n'est pas efficace en terme de réduction de la consommation car ces classes ont une faible probabilité d'apparition.
Enfin, la plupart des techniques présentées ne prennent pas toujours en compte le surcoût en consommation dû aux codecs. C'est pourquoi, les résultats présentés dans [9] montrent que le surcoût en consommation des codecs est souvent supérieur à la réduction de consommation apportée sur le bus.
Basé sur ces observations présentées dans [7], de nouvelles pistes d'optimisation pour le délai et la consommation des interconnexions peuvent être mise en avant :

- Ne pas se focaliser uniquement sur les dernières classes de transition de la classification en temporel.
- Porter l'optimisation en consommation sur les lignes présentant la plus forte activité (i.e. les lignes de poids faible) car ce sont les plus consommatrices.
- Essayer d'éliminer les transitions descendantes, ou coder les données telles que ces transitions présentent la plus faible capacité possible.
- Designer les codecs tel que le surcoût en consommation soit le plus faible possible, et donc trouver des techniques d'encodage très simples.

Grâce à ces observations, nous avons mis au point de nouvelles techniques d'encodage permettant d'obtenir des gains en consommation sur le bus avoisinant 13 % pour des longueurs de bus faibles ($\leq$ 3mm), dans les technologies actuelles (90, 65 nm), ceci en incluant dans les mesures le surcoût en consommation apporté par les codecs.

4. BIBLIOGRAPHIE

[1] N. Magen, A. Kolodny, U. Weiser, and N. Shamir, "Interconnect-power dissipation in a microprocessor," in *the Proceedings of the 2004 international workshop on System level interconnect prediction (SLIP)*, Paris, France, 2004, pp. 7–13.

[2] R. Ho, K. Mai, and M. Horowitz, "The future of wires," *Proceedings . IEEE*, vol. 89, no. 4, pp. 490–504, April 2001.

[3] L. Benini, G. D. Micheli, E. Macii, D. Sciuto, and C. Silvano, "Asymptotic zero-transition activity encoding for address busses in low-power microprocessor-based systems," in *the Proceedings of the 7th Great Lakes Symposium on VLSI (GLS)*, Urbana, USA, 1997, p. 77.

[4] J. Philippe, S. Pillement, and O. Sentieys, "Area efficient temporal coding schemes reducing crosstalk effects," in *the Proceedings of the International Symposium on Quality Electronic Design (ISQED)*, San Jose, USA, 2006, pp. 334–339.

[5] M. Stan and W. P. Burleson, "Bus-invert coding for low-power i/o," in *IEEE Trans. on Very Large Scale Integration Systems*, vol. 3, no. 1, 1995, pp. 49–58.

[6] C. Su, C. Y. Tsu, and A. M. Despain, "Saving power in the control path of embedded processors," *Design & Test of Computers, IEEE*, vol. 11, pp. 24–31, 1994.

[7] A. Courtay, O. Sentieys, J. Laurent, and N. Julien, "High-level interconnect delay and power estimation," *Journal of Low Power Electronics*, pp. 21–33, 2008.

[8] P. Landman and J. Rabaey, "Architectural power analysis : the dual bit type method," in *IEEE Trans. on Very Large Scale Integration Systems*, vol. 3, no. 2, 1995, pp. 173–187.

[9] C. Kretzschmar, A. Nieuwland, and D. Muller, "Why transition coding for power minimization of on-chip buses does not work," in *the Proceedings of the Conference on Design, Automation and Test in Europe (DATE)*, 2004, pp. 10 512–10 517.