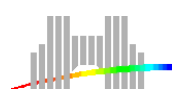
Un exercice d'optimisation sous contrainte

Paul Feautrier

ENS de Lyon
Paul.Feautrier@ens-lyon.fr

15 mai 2007



Hypothèses, I

- Un système composé de n blocs. t_k durée d'exécution de B_k.
- Chaque bloc reçoit des données de ses prédécesseurs et envoie ses résultats à des successeurs.
- Le graphe de précédence est acyclique. On suppose les B_k numérotés dans l'ordre du tri topologique.
- Notations $B_i \in \mathrm{pred}(B_j)$, etc.
- Date de lancement de B_k :

$$T_k = \max_{i \in \mathrm{pred}(B_k)} T_i + t_i.$$

Hypothèses, II

- Chaque bloc B_k doit exécuter un nombre fixé d'avance de cycles de calcul, n_k.
- La tension d'alimentation de B_k est V_k. En conséquence, l'énergie dissipée à chaque cycle est $C_k V_k^2$, où C_k est une capacité équivalente, qui reflète la taille de B_k et son taux d'activité.
- Au total, l'énergie consommée par B_k est $e_k V_k^2$, avec $e_k = C_k n_k$, et l'énergie consommée par le système est :

$$E = \sum_{k=1}^{n} e_k V_k^2.$$

Hypothèses, III

- On trouve dans la littérature[1] la formule suivante pour le temps de traversée d'un inverseur CMOS la formule suivante :

$$\tau = \frac{C}{h(V_{DD} - V_T)}\left[\frac{2V_T}{V_{DD} - V_T} + \ln\left(4\frac{V_{DD} - V_T}{V_{DD}} - 1\right)\right]$$

- Si $V_{DD} \gg V_T$, ceci se simplifie en

$$\tau = \frac{C \ln 3}{kV_{DD}}.$$

- Au total, si on admet que τ est une estimation du temps de cycle, le calcul effectué par B_k prent le temps :

$$t_k = n_k \frac{C \ln 3}{hV_k} = \theta_k / V_k.$$

[1] Sung-Mo Kang et Yusuf Leblebici, *CMOS Integrated Circuits Analysis and Design*, Mac-Graw Hill, 2003. h est la transconductance du CMOS.

En résumé ...

- Minimiser :

$$E = \sum_{k=1}^{n} e_k V_k^2,$$

- Sous la contrainte $T_n \leq T$, avec

$$T_k = \max_{i \in \mathrm{pred}(B_k)} T_i + \theta_i / V_i.$$

- Il est évident que les solutions où la contrainte $T_n \leq T$ est saturée dominent les autres solutions.

Le cas séquentiel

- Chaque bloc n'a qu'un seul prédécesseur et qu'un seul successeur (au plus).

$$\begin{aligned} & \min \quad \sum_k e_k V_k^2 \\ T \quad = \quad & \sum_k \theta_k / V_k . \end{aligned}$$

Résolution

- Les variations admissibles dE_k satisfaire la contraite $dT = 0$, soit :

$$\sum_k \theta_k dV_k / V_k^2 = 0,$$

- et doivent laisser la fonction objectif stationnaire :

$$\sum e_k V_k dv_k = 0,$$

- ce qui est réalisé si $e_k V_k^3 = \lambda^3 \theta_k$, soit $V_k = \lambda(\theta_k / e_k)^{1/3}$.

Résolution, suite

- On détermine λ à partir de la contrainte temporelle :

$$T = \sum_k \theta_k / V_k = \lambda^{-1} \sum_k \theta_k^{2/3} e_k^{1/3}$$

$$\lambda = \frac{\sum_k \theta_k^{2/3} e_k^{1/3}}{T}$$

- et l'énergie minimum :

$$E = \lambda^2 \sum_k \theta_k^{2/3} e_k^{1/3} = \frac{(\sum_k \theta_k^{2/3} e_k^{1/3})^3}{T^2}.$$

- Vérification par les dimensions :
  - $[e_k] : [E]/[V]^2$, $[\theta_k] : [T][V]$,
  - $[e_k^{1/3} \theta_k^{2/3}] : [E]^{1/3}[T]^{2/3}$,
  - $[E] = [E][T]^2/([T]^2) = [E]$.

Le cas parallèle

- Toutes les branches doivent prendre le temps maximum :

$$T = \theta_k / V_k, \quad V_k = \theta_k / T.$$

$$E = \frac{\sum_k e_k \theta_k^2}{T^2}.$$

- Comme dans le cas précédent, l'énergie varie en raison inverse du carré du temps imparti. Ceci permet une résolution complète d'un circuit série parallèle.

Parallélisme et économie d'énergie

- Soit un bloc caractérisé par les coefficients e et θ dont la latence imposée est T
- On le remplace par deux blocs identiques. On suppose que la tâche à éxécuter peut se paralléliser exactement. Si donc elle demandait n cycles, chaque bloc en exécute maintenant $n/2$. Le coefficient e est donc divisé par 2.
- De la même façon, le coefficient θ est divisé par deux.
- Enfin, la latence imposée ne change pas.
- Au total, l'énergie optimale : $E = 2\frac{e/2.\theta^2/4}{T^2}$ est divisée par 4.

Mise en série

- On connecte en série deux sous-circuits de latence T_1 et T_2 et dont les consommations optimales sont ϵ_1^3/T_1^2 et ϵ_2^3/T_2^2.
- On veut que la latence du circuit complet soit $T = T_1 + T_2$.
- A l'optimum $T_1/\epsilon_1 = T_2/\epsilon_2$, et l'énergie optimale est $E = (\epsilon_1 + \epsilon_2)^3/T^2$.

Le cas général

- Dans le cas d'un seul bloc caractérisé par e et θ, l'énergie est donnée par $E = \frac{e\theta^2}{T^2}$. On pose $\epsilon^3 = e\theta^2$.
- Soit deux circuits caractérisés par les lois de consommation

$$E_1 = \epsilon_1^3/T_1^2, \quad E_2 = \epsilon_2^3/T_2^2.$$

- Si on les connecte en parallèle, la consommation optimale est obtenue pour $T_1 = T_2 = T$, et vaut :

$$E = (\epsilon_1^3 + \epsilon_2^3)/T^2.$$

- Si on les connecte en série, la consommation optimale est obtenue pour $T_1/T_2 = \epsilon_1/\epsilon_2$, et vaut :

$$E = (\epsilon_1 + \epsilon_2)^3/T^2.$$

- Dans les deux cas, la *forme* de la loi de consommation n'est pas modifiée. Il est donc possible d'itérer.