IMPLANTATION MATÉRIELLE D'UNE MÉTHODE DE CONTRÔLE DES ERREURS D'ARRONDI DE CALCUL

Roselyne CHOTIN Habib MEHREZ

Laboratoire LIP6/ASIM, Université Pierre et Marie CURIE
4, place Jussieu, 75252 Paris Cedex 05
Roselyne.Chotin@lip6.fr
Habib.Mehrez@lip6.fr

RÉSUMÉ : Il est reconnu qu'utiliser les nombres flottants dans un programme de calcul pose des problèmes de précision de ces calculs, puisqu'en machine un nombre est représenté sur un nombre fini de bits, alors qu'en arithmétique exacte un nombre peut avoir une infinité de décimales. On voudrait donc pouvoir contrôler et estimer cette perte de précision et c'est ce que fait la méthode CESTAC (Contrôle et Estimation STochastique des Arrondis de Calculs). L'implantation logicielle de cette méthode requiert énormément de temps de calcul. Cet article présente une alternative matérielle qui devrait permettre d'augmenter significativement les performances.

I – INTRODUCTION

L'utilisation des nombres flottants pour le calcul scientifique est altérée par le fait que chaque opération peut générer une erreur d'arrondi qui se répercute ensuite sur les calculs suivants. Cette erreur est telle que parfois le résultat obtenu est totalement différent du résultat escompté. Le but de l'approche stochastique et de la méthode CESTAC est d'estimer cette erreur, et ainsi de pouvoir contrôler la validité des résultats fournis par la machine.
Une version logicielle de la méthode CESTAC existe, mais est très coûteuse en temps de calcul. La version matérielle vise donc à en améliorer les performances.
Cet article présente l'implantation matérielle de la méthode CESTAC. La deuxième partie de ce document sera consacré à une brève présentation des bases de l'arithmétique stochastique et de sa version logicielle. Les choix architecturaux seront détaillés dans la troisième partie. Nous présenterons dans la quatrième partie les performances actuelles du matériel développé. Enfin la conclusion sera donnée dans la cinquième partie.

II – L'ARITHMÉTIQUE STOCHASTIQUE

La méthode CESTAC a été développée par M. La Porte et J. Vignes [1] [2][3]. L'idée principale de la méthode consiste à exécuter plusieurs fois le même programme en propageant différemment les erreurs d'arrondi. Le résultat significatif sera la partie commune des différents résultats ainsi obtenus.
Un nombre stochastique a trois composantes flottantes. Effectuer une opération stochastique revient donc à effectuer cette opération sur chacune des composantes des opérandes stochastiques, le résultat étant arrondi aléatoire-

ment vers $\pm\infty$. L'exemple suivant présente une séquence d'opérations utilisant l'arithmétique stochastique.

$$\left(\frac{2}{3}+\frac{1}{3}\right)\times 3 = \left(\left|\begin{array}{c}2.00000\\2.00000\\2.00000\end{array}\right| \div \left|\begin{array}{c}3.00000\\3.00000\\3.00000\end{array}\right| + \left|\begin{array}{c}1.00000\\1.00000\\1.00000\end{array}\right| \div \left|\begin{array}{c}3.00000\\3.00000\\3.00000\end{array}\right|\right) \times \left|\begin{array}{c}3.00000\\3.00000\\3.00000\end{array}\right|$$

$$= \left(\left|\begin{array}{c}0.66666 \rightarrow -\infty\\0.66666 \rightarrow -\infty\\0.66667 \rightarrow +\infty\end{array}\right| + \left|\begin{array}{c}0.33333 \rightarrow -\infty\\0.33334 \rightarrow +\infty\\0.33333 \rightarrow -\infty\end{array}\right|\right) \times \left|\begin{array}{c}3.00000\\3.00000\\3.00000\end{array}\right|$$

$$= \left|\begin{array}{c}0.99999 \rightarrow +\infty\\1.00000 \rightarrow -\infty\\1.00000 \rightarrow +\infty\end{array}\right| \times \left|\begin{array}{c}3.00000\\3.00000\\3.00000\end{array}\right| = \left|\begin{array}{c}2.99997 \rightarrow +\infty\\3.00000 \rightarrow -\infty\\3.00000 \rightarrow +\infty\end{array}\right|$$

II–1 – L'estimation de la précision

Le résultat final du calcul est la moyenne des différents résultats calculés par la méthode, soit $R = \frac{1}{N}\sum_{i=1}^{N} R_i$. Le nombre de chiffres significatifs de R s'exprime alors par :

$$C_R = \log_{10} \frac{\sqrt{N}.|R|}{s.\tau_\beta} \; avec \; s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(R_i - R)^2 \tag{1}$$

En pratique : Pour $N = 2,\ \tau_\beta = 12.706$ et pour $N = 3,\ \tau_\beta = 4.303$
Avec l'exemple précédent, le résultat final vaut 2.99999 et le nombre de chiffres significatifs est 4.84.

II–2 – Le zéro informatique

Comme tout résultat d'un calcul flottant est entaché d'erreurs d'arrondi, il se peut que ces erreurs se répercutent sur un résultat censé être nul. La notion de zéro informatique a donc été introduite [4]. Un résultat informatique R est un zéro informatique (noté $\underline{0}$) si une des conditions suivantes est vraie :

1. $\forall\, i,\ R_i = 0,$
2. $C_R \leq 0.$

II–3 – Le logiciel CADNA

Une version logicielle de la méthode a été développée sous la forme d'une librairie nommée CADNA (*Control of Accuracy and Debugging for Numerical Applications*). Cette librairie permet de pouvoir estimer l'impact des erreurs d'arrondi dans tout résultat de programmes scientifiques et surtout d'avoir un véritable débogage numérique.

II–3.1 – Exemple d'utilisation

Un programme calcule les racines de l'équation du second degré $0.3x^2 - 2.1x + 3.675 = 0$. Cette équation admet pour racine double $x = 3.5$. Lorsqu'on exécute ce programme sur machine, le discriminant calculé vaut -3.8146972E-06 et du coup il y a deux racines complexes conjuguées. L'utilisation de la librairie CADNA permet de détecter que le discriminant est un zéro informatique et par conséquent que l'équation admet une racine double réelle.

III – IMPLANTATION MATÉRIELLE

L'implantation matérielle de la méthode CESTAC requiert de calculer le nombre de bits significatifs d'un résultat à trois composantes flottantes et de pouvoir détecter les zéros informatiques.

III–1 – Calcul du nombre de bits significatifs

Le nombre de chiffres significatifs est donné par la formule (1). Cette formule est trop complexe pour pouvoir être implantée directement en matériel et il a donc fallu la simplifier, ce qui revient à :

- On calcul $d_1 = |R_1 - R_2|$, $d_2 = |R_1 - R_3|$, $d_3 = |R_3 - R_2|$
- Pour chaque d_i on cherche la position du premier bit à 1 (p_i)
- Le nombre de bits significatifs est $min(p_1, p_2, p_3)$

Cette nouvelle méthode de calcul a été validée par comparaison avec la librairie CADNA. Elle est facile à implanter en matériel au moyen de trois opérateurs qui calculent la valeur absolue de la différence, trois encodeurs de priorité pour rechercher la position du premier bit à 1, trois comparateurs et quelques portes logiques pour récupérer le minimum. L'architecture de cet opérateur est présentée par la figure 1.

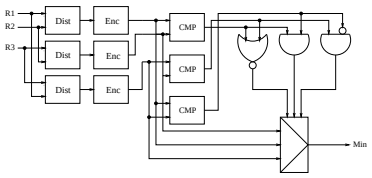

FIG. 1 – Calcul du nombre de bits significatifs

III–2 – Détection des zéros informatiques

Pour détecter un zéro informatique, il faut vérifier une des conditions suivantes :

1. $\forall\, i,\ R_i = 0$,
2. $C_R \leq 0$.

En matériel, on teste si les trois résultats sont nuls ou si le nombre de bits significatifs vaut zéro et on retourne le résultat de ce test. Ceci se fait facilement avec un peu de logique combinatoire, comme le montre la figure 2.

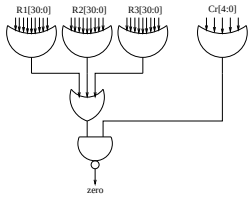

FIG. 2 – Détection du zéro informatique

III–3 – Assemblage du tout

La méthode CESTAC s'utilise avec du calcul flottant. Il va donc falloir intégrer les éléments décrits précédemment à une unité flottante. On a vu qu'effectuer une opération stochastique revenait à exécuter l'opération sur chacune des trois composantes des opérandes stochastiques. L'architecture de l'opérateur complet est présentée par la figure 3. Le bloc CESTAC va stocker chacun des trois résultats puis calculer le nombre de bits significatifs et détecter si on est en présence d'un zéro informatique.

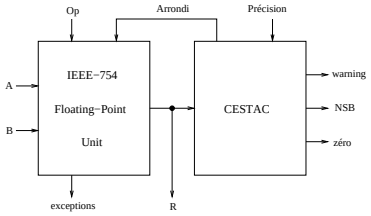

FIG. 3 – L'unité flottante stochastique

IV – PERFORMANCES

Le tableau 1 donne les performances de l'unité flottante stochastique actuelle, comprenant un additionneur, un soustracteur, un comparateur, des convertisseurs entier/flottant. La bibliothèque de cellules utilisée est ALLIANCE/SXLIB [5]. Elle a été placée et routée avec CADENCE/SILICON ENSEMBLE et l'analyse temporelle a été réalisée avec AVERTEC/TAS[1].

Taille (bits)	Techno (μm)	Surface (mm^2)	Fréquence (MHz)
32	0.35	1.51	59
64	0.35	2.19	52
32	0.25	0.18	187
64	0.25	0.25	167

TAB. 1 – Performances de l'unité flottante stochastique

V – CONCLUSION

Nous avons développé une unité flottante stochastique effectuant l'addition, la soustraction, la comparaison et les conversions entier/flottant. Avec une technologie de 0.25 μm, cette unité 32 bits représente une surface de 0.18 mm^2 et fonctionne à la fréquence de 187 MHz. Maintenant pour utiliser cette unité nous avons à l'intégrer dans un cœur de processeur et envisageons de l'implanter sur un FPGA. D'autres opérateurs tels que la multiplication, la division et la racine-carrée sont en cours de développement. Une étude comparative matériel/logiciel en terme de performances est également en cours.

RÉFÉRENCES

[1] M. Pichat and J. Vignes, *Ingénierie du contrôle de la précision des calculs sur ordinateur*. technip ed., 1993.

[2] J. Vignes, "Review on stochastic approach to round-off error analysis and its applications," *Math. Comp. Simul.*, vol. 30, pp. 481–491, 1988.

[3] J. Vignes and M. La Porte, "Error analysis in computing," in *Information Processing 74*, (North-Holland), 1974.

[4] J. Vignes, "Zéro mathématique et zéro informatique," in *La vie des Sciences, C.R. Acad. Sci.*, no. 1 in 4, pp. 1–13, Jan. 1987.

[5] A. Greiner and al ALLIANCE, "A complet set of CAD Tools for teaching VLSI Design," *Third EuroChip Workshop*, 1992. http ://www-asim.lip6.fr/alliance.

[1] http ://www.avertec.com