Arrondi correct en demi-précision (stage L3)

Encadrant. Paul Zimmermann (INRIA Nancy - Grand Est).

Contexte. Le standard IEEE 754 définit ce qu'on appelle l'« arrondi correct ». Étant donnés une fonction mathématique f et un nombre flottant x, l'arrondi correct de $f(x)$ est le nombre flottant y le plus proche de $f(x)$ selon le mode d'arrondi donné (au plus proche, vers zéro, vers $-\infty$ ou vers $+\infty$). Si le standard IEEE 754 impose l'arrondi correct pour les quatre opérations arithmétiques de base (addition, soustraction, multiplication, division) et la racine carrée, il ne fait que le recommander pour les fonctions mathématiques (exp, sin, pow). Les bibliothèques mathématiques couramment utilisées ne garantissent pas l'arrondi correct [1]. Le projet CORE-MATH [4, 2] vise à produire des implantations avec arrondi correct, en vue d'une intégration dans ces bibliothèques mathématiques.

Objectif du stage. L'objectif du stage est d'implanter des fonctions avec arrondi correct pour le format `fp16` (demi-précision). Ce format est codé sur 16 bits : 1 bit de signe, 5 bits d'exposant, et 11 bits de mantisse (dont 1 bit implicite). Ce format est supporté par GCC depuis la version 12, sous le nom `_Float16`. Une première étape sera de vérifier que les opérations arithmétiques de base (addition, soustraction, multiplication, division) sont correctement arrondies en `fp16`. Pour cela on pourra réaliser des tests exhaustifs en comparant aux valeurs calculées par GNU MPFR pour chacun des modes d'arrondi de IEEE 754. En effet, comme il y a 2^{16} valeurs possibles en demi-précision, pour une opération de deux variables, il y a 2^{32} valeurs à tester, ce qui est atteignable. On implantera ensuite les fonctions mathématiques usuelles (sin, exp, log, ...) avec arrondi correct. Pour cela, on pourra comparer trois approches : (i) utiliser les algorithmes classiques de [3] ; (ii) utiliser le code simple précision de CORE-MATH ; (iii) tabuler les valeurs de $f(x)$ pour certaines fonctions où on peut se ramener à de petites tables. Ces implantations en C seront effectuées dans le cadre de CORE-MATH (`https://core-math.gitlabpages.inria.fr/`).

Prérequis. Ce stage nécessite de solides connaissances mathématiques, ainsi qu'une bonne maîtrise du langage C.

Références

[1] Gladman, B., Innocente, V., Mather, J., and Zimmermann, P. Accuracy of mathematical functions in single, double, extended double and quadruple precision. `https://members.loria.fr/PZimmermann/papers/accuracy.pdf`, 2024. Version of August, 26 pages.

[2] Hubrecht, T., Jeannerod, C.-P., Zimmermann, P., Rideau, L., and Théry, L. Towards a correctly-rounded and fast power function in binary64 arithmetic. This is the extended version of an article published in the proceedings of ARITH 2023, available at `https://inria.hal.science/hal-04159652`, Feb. 2024.

[3] Markstein, P. *IA-64 and Elementary Functions*. Hewlett-Packard Professional Books, 2000.

[4] Sibidanov, A., Zimmermann, P., and Glondu, S. The CORE-MATH Project. In *ARITH 2022 - 29th IEEE Symposium on Computer Arithmetic* (virtual, France, Sept. 2022). `https://hal.inria.fr/hal-03721525`.