

# Arrondi impair et sans reproche

## Émulation du FMA et sommes correctement arrondies

Guillaume Melquiond

Laboratoire de l'Informatique du Parallélisme  
Arénaire, LIP, CNRS-ENSL-INRIA-UCBL

2007-01-25

# Plan

- 1 Introduction
- 2 Arrondi impair
- 3 Sommutation avec arrondi correct
- 4 Émulation du Fused-Multiply-and-Add
- 5 Conclusion

# L'arithmétique flottante vue de loin

## Définition

Format :  $B = (p, E)$  (53, 1074 pour la double précision)

Nombre flottants :

$$\mathbb{F}_B = \{ m \cdot 2^e \mid m, e \in \mathbb{Z}, |m| < 2^p \wedge e \geq -E \}.$$

# L'arithmétique flottante vue de loin

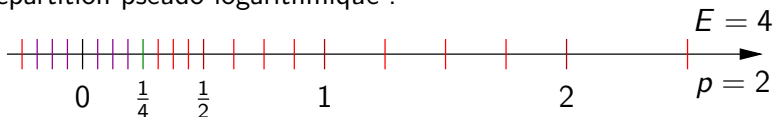
## Définition

Format :  $B = (p, E)$  (53, 1074 pour la double précision)

Nombre flottants :

$$\mathbb{F}_B = \{ m \cdot 2^e \mid m, e \in \mathbb{Z}, |m| < 2^p \wedge e \geq -E \}.$$

Répartition pseudo-logarithmique :



# L'arithmétique flottante vue de loin

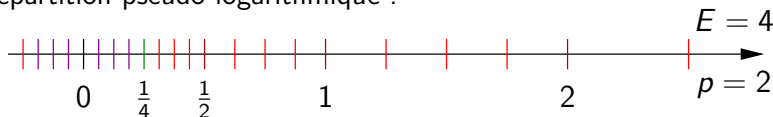
## Définition

Format :  $B = (p, E)$  (53, 1074 pour la double précision)

Nombre flottants :

$$\mathbb{F}_B = \{ m \cdot 2^e \mid m, e \in \mathbb{Z}, |m| < 2^p \wedge e \geq -E \}.$$

Répartition pseudo-logarithmique :



Cas à problèmes : puissance de deux et **nombre sous-normal**.

Notation :  $\mu_B = 2^{p-E}$  est le plus petit flottant positif **normal**.

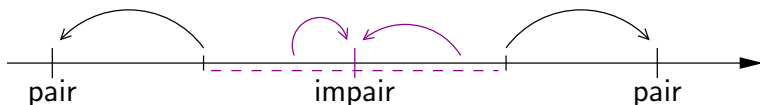
# Arrondi et double arrondi

- 1 Arrondi : fonction  $\mathbb{R} \rightarrow \mathbb{F}$ , identité sur  $\mathbb{F}$ , croissante sur  $\mathbb{R}$ .

# Arrondi et double arrondi

- 1 Arrondi : **fonction**  $\mathbb{R} \rightarrow \mathbb{F}$ , **identité** sur  $\mathbb{F}$ , **croissante** sur  $\mathbb{R}$ .

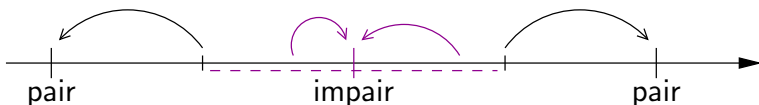
Arrondi **au plus près pair** :



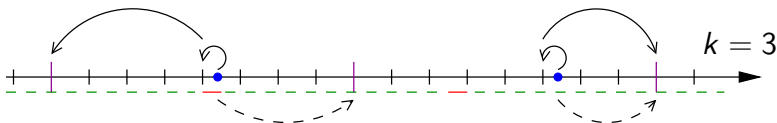
# Arrondi et double arrondi

- ① Arrondi : **fonction**  $\mathbb{R} \rightarrow \mathbb{F}$ , **identité** sur  $\mathbb{F}$ , **croissante** sur  $\mathbb{R}$ .

Arrondi **au plus près pair** :



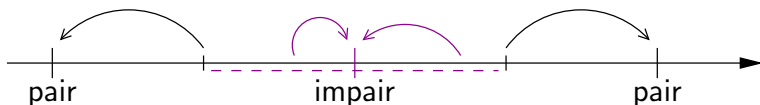
- ② Double arrondi :  $\circ_p(\circ_{p+k}(x)) = \circ_p(x)$  ?



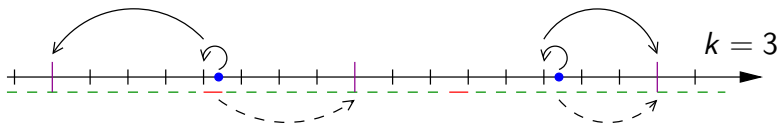
# Arrondi et double arrondi

- ① Arrondi : **fonction**  $\mathbb{R} \rightarrow \mathbb{F}$ , **identité** sur  $\mathbb{F}$ , **croissante** sur  $\mathbb{R}$ .

Arrondi **au plus près pair** :



- ② Double arrondi :  $\circ_p(\circ_{p+k}(x)) = \circ_p(x)$  ?



Probabilité d'**échec** :  $2^{-k-1}$

( $\approx 1/4000$  sur processeur x86)

# Arrondi impair

Historique :

# Arrondi impair

Historique :

- 1 Von Neumann, EDVAC, 1945.

# Arrondi impair

## Historique :

- 1 Von Neumann, EDVAC, 1945.
- 2 Annexe Z de la future norme IEEE-754R, 2005.

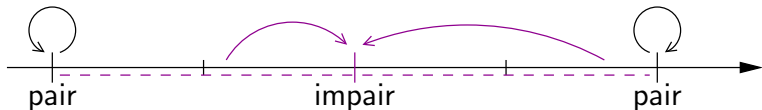
# Arrondi impair

Historique :

- 1 Von Neumann, EDVAC, 1945.
- 2 Annexe Z de la future norme IEEE-754R, 2005.

## Définition (Arrondi impair)

$$\square(x) = \begin{cases} x & \text{si } x \in \mathbb{F}, \\ \triangle(x) & \text{si } \triangle(x) \text{ a une mantisse impaire,} \\ \nabla(x) & \text{sinon.} \end{cases}$$



# Arrondi impair et double arrondi

## Théorème (To\_Odd\_Even\_Is\_Even)

Soient  $B = (p, E)$  et  $B_{ext} = (p + k, E_{ext})$  deux formats flottants avec  $p \geq 2$ ,  $k \geq 2$  et  $E_{ext} \geq 2 + E$ ,

$$\forall x \in \mathbb{R}, \circ(\square_{ext}(x)) = \circ(x).$$

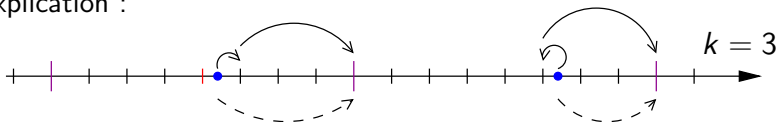
# Arrondi impair et double arrondi

## Théorème (To\_Odd\_Even\_Is\_Even)

Soient  $B = (p, E)$  et  $B_{ext} = (p + k, E_{ext})$  deux formats flottants avec  $p \geq 2$ ,  $k \geq 2$  et  $E_{ext} \geq 2 + E$ ,

$$\forall x \in \mathbb{R}, \circ(\square_{ext}(x)) = \circ(x).$$

Explication :



Le **milieu** de deux flottants consécutifs dans le format  $B$  est un flottant **pair** dans le format  $B_{ext}$ , donc repoussant.

# Double arrondi et preuve formelle

Théorèmes prouvés en **Coq**. Avantages :

- 1 Un assistant de preuve permet de s'assurer qu'aucun cas particulier n'a été oublié.

# Double arrondi et preuve formelle

Théorèmes prouvés en **Coq**. Avantages :

- 1 Un assistant de preuve permet de s'assurer qu'aucun cas particulier n'a été oublié.
- 2 Il oblige (parfois) à chercher des preuves élégantes.  
⇒ Nouveaux algorithmes.

# Double arrondi et preuve formelle

Théorèmes prouvés en **Coq**. Avantages :

- 1 Un assistant de preuve permet de s'assurer qu'aucun cas particulier n'a été oublié.
- 2 Il oblige (parfois) à chercher des preuves élégantes.  
⇒ Nouveaux algorithmes.

## Exemple

Q : Comment prouver que  $v \in \mathbb{F}$  vaut  $\circ_p(x)$  ?

# Double arrondi et preuve formelle

Théorèmes prouvés en **Coq**. Avantages :

- 1 Un assistant de preuve permet de s'assurer qu'aucun cas particulier n'a été oublié.
- 2 Il oblige (parfois) à chercher des preuves élégantes.  
⇒ Nouveaux algorithmes.

## Exemple

Q : Comment prouver que  $v \in \mathbb{F}$  vaut  $\circ_p(x)$  ?

R : En prouvant que  $v$  vaut en fait  $\circ_p(\square_{p+k}(x))$ .

# Sommat

Soient  $f_1, \dots, f_n$  des flottants en précision  $p$ .

Notation :  $s_j = \sum_{i=1}^j f_i$ .

**Objectif** : calculer  $\circ_p(s_n)$  la somme correctement arrondie.

# Sommation – version 1.0

Soient  $f_1, \dots, f_n$  des flottants en précision  $p$ .

Notation :  $s_j = \sum_{i=1}^j f_i$ .

**Objectif** : calculer  $\circ_p(s_n)$  la somme correctement arrondie.

**Propriété** :  $\square_{p+k}(f_{i+1} + \square_{p+k}(s_i)) = \square_{p+k}(f_{i+1} + s_i)$ .

# Sommation – version 1.0

Soient  $f_1, \dots, f_n$  des flottants en précision  $p$ .

Notation :  $s_j = \sum_{i=1}^j f_i$ .

**Objectif** : calculer  $\circ_p(s_n)$  la somme correctement arrondie.

**Propriété** :  $\square_{p+k}(f_{i+1} + \square_{p+k}(s_i)) = \square_{p+k}(f_{i+1} + s_i)$ .

## Algorithme

$g_1 \leftarrow f_1$

pour  $i$  de 1 à  $n - 1$

$g_{i+1} \leftarrow \square_{p+k}(f_{i+1} + g_i)$

renvoie  $\circ_p(g_n)$

# Sommaton – version 2.0

Propriétés :

- $\square_p(f_{i+1} + \square_p(s_i)) = \square_p(f_{i+1} + s_i),$

# Sommaton – version 2.0

## Propriétés :

- $\square_p(f_{i+1} + \square_p(s_i)) = \square_p(f_{i+1} + s_i),$
- $\exists k, f_n + \square_p(s_{n-1}) = \square_{p+k}(f_n + s_{n-1}).$

# Sommmation – version 2.0

Propriétés :

- $\square_p(f_{i+1} + \square_p(s_i)) = \square_p(f_{i+1} + s_i),$
- $\exists k, f_n + \square_p(s_{n-1}) = \square_{p+k}(f_n + s_{n-1}).$

## Algorithme

```
 $g_1 \leftarrow f_1$   
pour  $i$  de 1 à  $n - 2$   
     $g_{i+1} \leftarrow \square_p(f_{i+1} + g_i)$   
renvoie  $\square_p(f_n + g_{n-1})$ 
```

La précision étendue n'est plus nécessaire.

# Théorèmes et contraintes

## Théorème (AddOddOdd2)

Soit  $x \in \mathbb{F}$  tel que  $|x| \geq 2 \cdot \mu$ . Soit  $z \in \mathbb{R}$ .

Si  $\frac{1}{2}$  est un flottant normal et si  $2 \cdot |z| \leq |x|$ ,

$$\square(x + \square(z)) = \square(x + z).$$

## Théorème (AddOddEven)

Soit  $x \in \mathbb{F}$  tel que  $|x| \geq 5 \cdot \mu$ . Soient  $z \in \mathbb{R}$  et  $y = \square(z)$ .

Si  $5 \cdot |y| \leq |x|$ ,

$$\circ(x + y) = \circ(x + z).$$

# Théorèmes et contraintes

## Théorème (AddOddOdd2)

Soit  $x \in \mathbb{F}$  tel que  $|x| \geq 2 \cdot \mu$ . Soit  $z \in \mathbb{R}$ .

Si  $\frac{1}{2}$  est un flottant normal et si  $2 \cdot |z| \leq |x|$ ,

$$\square(x + \square(z)) = \square(x + z).$$

## Théorème (AddOddEven)

Soit  $x \in \mathbb{F}$  tel que  $|x| \geq 5 \cdot \mu$ . Soient  $z \in \mathbb{R}$  et  $y = \square(z)$ .

Si  $5 \cdot |y| \leq |x|$ ,

$$\circ(x + y) = \circ(x + z).$$

**Contrainte** : les nombres flottants en entrée des algorithmes doivent être triés et suffisamment **espacés**.

# Expansions

## Définition

Expansion :

- $n$ -uplet de flottants  $(f_i)_i$  avec  $\forall i, |f_{i+1}| \geq 2^p \cdot |f_i|$ ;
- représente le réel  $\sum f_i$  avec une précision étendue.

Pseudo-expansion : léger recouvrement des termes autorisé.

# Expansions

## Définition

Expansion :

- $n$ -uplet de flottants  $(f_i)_i$  avec  $\forall i, |f_{i+1}| \geq 2^p \cdot |f_i|$ ;
- représente le réel  $\sum f_i$  avec une précision étendue.

Pseudo-expansion : léger recouvrement des termes autorisé.

**arrondir** une pseudo-expansion vers la précision courante



**sommer** les termes de la pseudo-expansion

# Expansions

## Définition

Expansion :

- $n$ -uplet de flottants  $(f_i)_i$  avec  $\forall i, |f_{i+1}| \geq 2^p \cdot |f_i|$ ;
- représente le réel  $\sum f_i$  avec une précision étendue.

Pseudo-expansion : léger recouvrement des termes autorisé.

**arrondir** une pseudo-expansion vers la précision courante



**sommer** les termes de la pseudo-expansion

Sommaton implantée dans **CRIbm** pour arrondir des **triples-doubles** (pseudo-expansions d'ordre 3) vers des flottants.

# Émulation du Fused-Multiply-and-Add

## Définition

$\text{fma} : (a, b, c) \in \mathbb{F}^3 \mapsto \circ(a \times b + c)$  (un seul arrondi !)

Comment **émuler** un FMA avec les opérations flottantes de base ?

# Émulation du Fused-Multiply-and-Add

## Définition

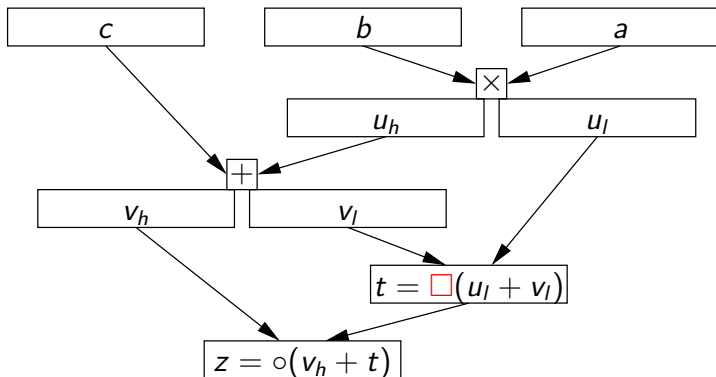
$$\text{fma} : (a, b, c) \in \mathbb{F}^3 \mapsto \circ(a \times b + c) \quad (\text{un seul arrondi !})$$

Comment **émuler** un FMA avec les opérations flottantes de base ?

Remarque : pour  $x, y \in \mathbb{F}$ , on sait calculer les flottants

- $(x + y) - \circ(x + y)$ , (Knuth)
- $(x \times y) - \circ(x \times y)$ . (Dekker)

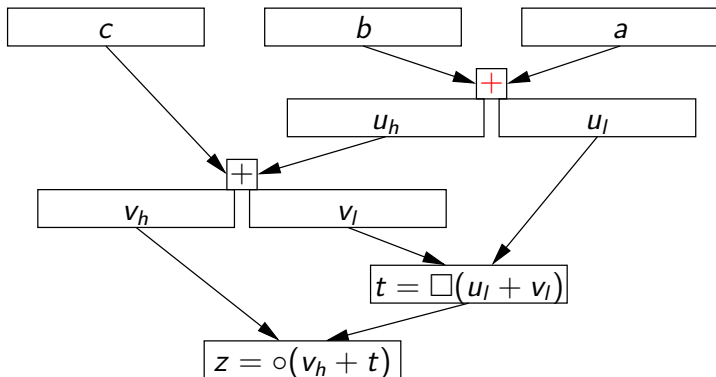
# Émulation du Fused-Multiply-and-Add



## Théorème (FmaEmul)

Si  $u_l$  est représentable et si  $p \geq 5$ , alors  $z = \circ(a \times b + c)$ .

## Variante : somme de trois flottants



## Théorème

Si  $p \geq 5$ , alors  $z = \circ(a + b + c)$ .

# Conclusion

L'arrondi impair est :

- similaire à l'arrondi vers zéro,

# Conclusion

L'arrondi impair est :

- similaire à l'arrondi vers zéro,
- moins précis que l'arrondi au plus près,

# Conclusion

L'arrondi impair est :

- similaire à l'arrondi vers zéro,
- moins précis que l'arrondi au plus près,
- mieux adapté pour les sommes intermédiaires.

# Conclusion

L'arrondi impair est :

- similaire à l'arrondi vers zéro,
- moins précis que l'arrondi au plus près,
- mieux adapté pour les sommes intermédiaires.

Les assistants de preuve sont pénibles, mais c'est pour notre bien.

# Questions ?

- Travail réalisé en collaboration avec [Sylvie Boldo](#) (projet ProVal, INRIA Futurs, PCRI).
- Développement formel en Coq disponible sur <http://lipforge.ens-lyon.fr/www/pff/>