RAIM 2013, Paris

Pourquoi faire simple quand on peut faire compliqué: calculer la racine carrée du carré

Sylvie Boldo

Équipe-projet Toccata

19 novembre 2013

Plan

1. Pourquoi ?
2. $|x| = \sqrt{x^2}$
3. $|x| \neq \sqrt{x^2}$

Pourquoi $\sqrt{x^2}$?

Pourquoi $\sqrt{x^2}$?

C'est avec ce problème que Jean-Michel Muller

Pourquoi $\sqrt{x^2}$?

C'est avec ce problème que Jean-Michel Muller

Pourquoi $\sqrt{x^2}$?

C'est avec ce problème que Jean-Michel Muller

m'a attirée vers l'arithmétique flottante.

Pourquoi $\sqrt{x^2}$?

Au millénaire précédent, il m'a montré son tableau

$$-1 \leq \frac{x}{\sqrt{x^2+y^2}} \leq 1$$

et m'a raconté les problèmes d'arrondis.

Pourquoi $\sqrt{x^2}$?

Or, le pire cas (la valeur la plus grande) pour $\frac{x}{\sqrt{x^2+y^2}}$ est pour $y = 0$, soit

$$\frac{x}{\sqrt{x^2}}.$$

Pourquoi $\sqrt{x^2}$?

Or, le pire cas (la valeur la plus grande) pour $\frac{x}{\sqrt{x^2+y^2}}$ est pour $y = 0$, soit

$$\frac{x}{\sqrt{x^2}}.$$

En fait,

$$\circ\left(\frac{x}{\circ\left(\sqrt{\circ\left(x^2\right)}\right)}\right).$$

État de l'art

Au fond d'un article sur son site web, Kahan dit :

Mathematics written in the sand, 1983 ?

[...] must be surprised to discover, on binary and quaternary machines but not on those with larger radix, that despite roundoff $\sqrt{x^2} = |x|$ *for all* x *unless* x^2 *over/underflows. These surprises can be confirmed first by experiment, then by simple proofs.*

Et donc ?

- Pour toute base et toute précision (si possible)...

Et donc ?

- Pour toute base et toute précision (si possible)...
- J'ai voulu prouver formellement les cas où $\sqrt{x^2} = |x|$.

Et donc ?

- Pour toute base et toute précision (si possible)...
- J'ai voulu prouver formellement les cas où $\sqrt{x^2} = |x|$.
- Dans les autres cas, j'ai regardé si on avait quand même $-1 \leq \frac{x}{\sqrt{x^2}} \leq 1$.

Et donc ?

- Pour toute base et toute précision (si possible)...
- J'ai voulu prouver formellement les cas où $\sqrt{x^2} = |x|$.
- Dans les autres cas, j'ai regardé si on avait quand même $-1 \leq \frac{x}{\sqrt{x^2}} \leq 1$.
- Preuve formelle (1 700 lines) utilisant Flocq.

Et donc ?

- Pour toute base et toute précision (si possible)...
- J'ai voulu prouver formellement les cas où $\sqrt{x^2} = |x|$.
- Dans les autres cas, j'ai regardé si on avait quand même $-1 \leq \frac{x}{\sqrt{x^2}} \leq 1$.
- Preuve formelle (1 700 lines) utilisant Flocq.
- On suppose une plage d'exposant infinie (ni overflow, ni underlow), ce qui correspond au format FLX de Flocq.

Plan

1. Pourquoi ?
2. $|x| = \sqrt{x^2}$
3. $|x| \neq \sqrt{x^2}$

Parfois, $x = \sqrt{x^2}$

On suppose un flottant $x > 0$ et la précision $p > 2$. Soit e tel que $\beta^{e-1} \leq x < \beta^e$. On réussit à prouver que $x = \sqrt{x^2}$

- quand $x = \sqrt{\beta}\beta^{e-1}$,

Parfois, $x = \sqrt{x^2}$

On suppose un flottant $x > 0$ et la précision $p > 2$. Soit e tel que $\beta^{e-1} \leq x < \beta^e$. On réussit à prouver que $x = \sqrt{x^2}$

- quand $x = \sqrt{\beta}\beta^{e-1}$,

  facile

Parfois, $x = \sqrt{x^2}$

On suppose un flottant $x > 0$ et la précision $p > 2$. Soit e tel que $\beta^{e-1} \leq x < \beta^e$. On réussit à prouver que $x = \sqrt{x^2}$

- quand $x = \sqrt{\beta}\beta^{e-1}$,

  facile
- quand $x < \sqrt{\beta}\beta^{e-1}$,

Parfois, $x = \sqrt{x^2}$

On suppose un flottant $x > 0$ et la précision $p > 2$. Soit e tel que $\beta^{e-1} \leq x < \beta^e$. On réussit à prouver que $x = \sqrt{x^2}$

- quand $x = \sqrt{\beta}\beta^{e-1}$,

  facile
- quand $x < \sqrt{\beta}\beta^{e-1}$,

  moyen : l'erreur du calcul de x^2 est alors petite, et divisée par 2 par la racine carrée.

Parfois, $x = \sqrt{x^2}$

On suppose un flottant $x > 0$ et la précision $p > 2$. Soit e tel que $\beta^{e-1} \leq x < \beta^e$. On réussit à prouver que $x = \sqrt{x^2}$

- quand $x = \sqrt{\beta}\beta^{e-1}$,

  facile
- quand $x < \sqrt{\beta}\beta^{e-1}$,

  moyen : l'erreur du calcul de x^2 est alors petite, et divisée par 2 par la racine carrée.
- quand $\beta \leq 4$,

Parfois, $x = \sqrt{x^2}$

On suppose un flottant $x > 0$ et la précision $p > 2$. Soit e tel que $\beta^{e-1} \leq x < \beta^e$. On réussit à prouver que $x = \sqrt{x^2}$

- quand $x = \sqrt{\beta}\beta^{e-1}$,

  facile
- quand $x < \sqrt{\beta}\beta^{e-1}$,

  moyen : l'erreur du calcul de x^2 est alors petite, et divisée par 2 par la racine carrée.
- quand $\beta \leq 4$,

  dur : ça va en bases 2 et 3, mais il faut regarder un peu précisément les plus petites valeurs $> \sqrt{\beta}\beta^{e-1}$ en base 4.

En base inférieure à 4

Theorem (round_flx_sqr_sqrt_exact)

Considérons un format flottant en base $\beta \leq 4$ sur $p > 2$ bits avec une plage d'exposants infinie et n'importe quels arrondis au plus proche, alors pour tout flottant x,

$$|x| = \circ\left(\sqrt{\circ\left(x^2\right)}\right).$$

En base inférieure à 4

Theorem (round_flx_sqr_sqrt_exact)

Considérons un format flottant en base $\beta \leq 4$ sur $p > 2$ bits avec une plage d'exposants infinie et n'importe quels arrondis au plus proche, alors pour tout flottant x,

$$|x| = \circ\left(\sqrt{\circ(x^2)}\right).$$

```
Notation format := (generic_format beta (FLX_exp prec)).
Variable choice1 choice2: Z -> bool.
Notation round_flx1:=(round beta (FLX_exp prec) (Znearest choice1)).
Notation round_flx2:=(round beta (FLX_exp prec) (Znearest choice2)).

Hypothesis pGt2: (2 < prec)%Z.

Theorem round_flx_sqr_sqrt_exact: forall x, format x ->
    (beta <= 4)%Z ->
        round_flx2(sqrt(round_flx1(x*x))) = Rabs x.
```

Plan

1. Pourquoi ?
2. $|x| = \sqrt{x^2}$
3. $|x| \neq \sqrt{x^2}$

Parfois, $|x| \neq \sqrt{x^2}$

- on trouve $x > \circ\left(\sqrt{\circ\left(x^2\right)}\right)$ (de plusieurs ulps).

Parfois, $|x| \neq \sqrt{x^2}$

- on trouve $x > \circ\left(\sqrt{\circ(x^2)}\right)$ (de plusieurs ulps).
  - Avec $\beta = 10$ et $p = 4$, si $x = 31.66$, alors $\circ\left(\sqrt{\circ(x^2)}\right) = 31.65$.

Parfois, $|x| \neq \sqrt{x^2}$

- on trouve $x > \circ\left(\sqrt{\circ(x^2)}\right)$ (de plusieurs ulps).
  - Avec $\beta = 10$ et $p = 4$, si $x = 31.66$, alors $\circ\left(\sqrt{\circ(x^2)}\right) = 31.65$.
  - Avec $\beta = 1000$ et $p = 2$, si $x = 31.662$, alors $\circ\left(\sqrt{\circ(x^2)}\right) = 31.654$.

Parfois, $|x| \neq \sqrt{x^2}$

- on trouve $x > \circ\left(\sqrt{\circ(x^2)}\right)$ (de plusieurs ulps).
  - Avec $\beta = 10$ et $p = 4$, si $x = 31.66$, alors $\circ\left(\sqrt{\circ(x^2)}\right) = 31.65$.
  - Avec $\beta = 1000$ et $p = 2$, si $x = 31.662$, alors $\circ\left(\sqrt{\circ(x^2)}\right) = 31.654$.
- mais la division renvoie quand même 1 !

Idée de la preuve (1/2)

- On suppose $x > 0$ et $x > \sqrt{\beta}\beta^{e-1}$, et on veut montrer
  - $\circ\left(\frac{x}{\circ\left(\sqrt{\circ(x^2)}\right)}\right) \leq 1.$

Idée de la preuve (1/2)

- On suppose $x > 0$ et $x > \sqrt{\beta}\beta^{e-1}$, et on veut montrer $\circ\left(\frac{x}{\circ\left(\sqrt{\circ(x^2)}\right)}\right) \leq 1$.
- Il suffit de montrer que $\frac{x}{\circ\left(\sqrt{\circ(x^2)}\right)} < 1 + \frac{\beta^{1-p}}{2}$.

Idée de la preuve (1/2)

- On suppose $x > 0$ et $x > \sqrt{\beta}\beta^{e-1}$, et on veut montrer $\circ\left(\frac{x}{\circ\left(\sqrt{\circ(x^2)}\right)}\right) \leq 1$.
- Il suffit de montrer que $\frac{x}{\circ\left(\sqrt{\circ(x^2)}\right)} < 1 + \frac{\beta^{1-p}}{2}$.
- On va considérer $x - k\,\mathrm{ulp}(x)$ et chercher des informations sur k.

Idée de la preuve (1/2)

- On suppose $x > 0$ et $x > \sqrt{\beta}\beta^{e-1}$, et on veut montrer $\circ\left(\frac{x}{\circ\left(\sqrt{\circ(x^2)}\right)}\right) \leq 1$.
- Il suffit de montrer que $\frac{x}{\circ\left(\sqrt{\circ(x^2)}\right)} < 1 + \frac{\beta^{1-p}}{2}$.
- On va considérer $x - k\,\mathrm{ulp}(x)$ et chercher des informations sur k.

  k est réputé positif et petit (pour que $x - k\,\mathrm{ulp}(x)$ soit un flottant ayant le même ulp que x).

Idée de la preuve (1/2)

- On suppose $x > 0$ et $x > \sqrt{\beta}\beta^{e-1}$, et on veut montrer $\circ\left(\frac{x}{\circ\left(\sqrt{\circ(x^2)}\right)}\right) \leq 1$.
- Il suffit de montrer que $\frac{x}{\circ\left(\sqrt{\circ(x^2)}\right)} < 1 + \frac{\beta^{1-p}}{2}$.
- On va considérer $x - k\,\mathrm{ulp}(x)$ et chercher des informations sur k.

  k est réputé positif et petit (pour que $x - k\,\mathrm{ulp}(x)$ soit un flottant ayant le même ulp que x).
- Si $2k\beta^{p-1}\mathrm{ulp}(x)\left(1 + \frac{\beta^{1-p}}{2}\right) < x$, alors $\circ\left(\frac{x}{x - k\,\mathrm{ulp}(x)}\right) \leq 1$.

Idée de la preuve (1/2)

- On suppose $x > 0$ et $x > \sqrt{\beta}\beta^{e-1}$, et on veut montrer $\circ\left(\frac{x}{\circ\left(\sqrt{\circ(x^2)}\right)}\right) \leq 1$.
- Il suffit de montrer que $\frac{x}{\circ\left(\sqrt{\circ(x^2)}\right)} < 1 + \frac{\beta^{1-p}}{2}$.
- On va considérer $x - k\,\mathrm{ulp}(x)$ et chercher des informations sur k.

  k est réputé positif et petit (pour que $x - k\,\mathrm{ulp}(x)$ soit un flottant ayant le même ulp que x).
- Si $2k\beta^{p-1}\mathrm{ulp}(x)\left(1 + \frac{\beta^{1-p}}{2}\right) < x$, alors $\circ\left(\frac{x}{x - k\,\mathrm{ulp}(x)}\right) \leq 1$.
- Si $\frac{\beta^e}{2} < (2k+1)x - \left(k + \frac{1}{2}\right)^2\beta^{e-p}$, alors $x - k\,\mathrm{ulp}(x) \leq \circ\left(\sqrt{\circ\left(x^2\right)}\right)$.

Idée de la preuve (2/2)

- On pose $k = \left\lceil \frac{x\beta^{1-e}}{1+\beta^{1-p}} \right\rceil - 1$.

Idée de la preuve (2/2)

- On pose $k = \left\lceil \frac{x\beta^{1-e}}{1+\beta^{1-p}} \right\rceil - 1$.
- On a $0 \leq k \leq \left\lceil \frac{\beta}{2} \right\rceil - 1$.

Idée de la preuve (2/2)

- On pose $k = \left\lceil \frac{x\beta^{1-e}}{1+\beta^{1-p}} \right\rceil - 1$.
- On a $0 \leq k \leq \left\lceil \frac{\beta}{2} \right\rceil - 1$.
- On a les 2 propriétés précédentes si

$$\sqrt{\beta} + \frac{\beta^{3-p}}{4} \leq \beta \frac{2 - \beta^{1-p}}{4 + 2\beta^{1-p}}.$$

Idée de la preuve (2/2)

- On pose $k = \left\lceil \frac{x\beta^{1-e}}{1+\beta^{1-p}} \right\rceil - 1$.
- On a $0 \leq k \leq \left\lceil \frac{\beta}{2} \right\rceil - 1$.
- On a les 2 propriétés précédentes si

$$\sqrt{\beta} + \frac{\beta^{3-p}}{4} \leq \beta \frac{2 - \beta^{1-p}}{4 + 2\beta^{1-p}}.$$

- OK si $\beta > 5$ ou $p > 3$.

En base supérieure à 5

Theorem (Muller)

Considérons un format flottant en base β sur $p > 2$ bits avec une plage d'exposants infinie et n'importe quels arrondis au plus proche.
On suppose également que si $\beta = 5$, alors $p > 3$.
Alors pour tout flottant x et tout y,

$$-1 \leq \circ\left(\frac{x}{\circ\left(\sqrt{\circ\left(\circ\left(x^2\right)+\circ\left(y^2\right)\right)}\right)}\right) \leq 1$$

En base supérieure à 5

```
Notation format := (generic_format radix (FLX_exp prec)).
Variable choice1 choice2 choice3 choice4 choice5 : Z -> bool.
Notation round_flx1 :=(round radix (FLX_exp prec) Znearest choice1)).
[...]

Hypothesis pGt2: (2 < prec)%Z.
Hypothesis pradix5: (beta=5)%Z -> (3 < prec)%Z.


Theorem Muller: forall x y:R, format x ->
   -1 <=  round_flx1 (x / round_flx2(
        sqrt (round_flx3(round_flx4(x*x)+round_flx5(y*y))))) <= 1.
```

Conclusion/Perspectives

- Muller avait raison.

Conclusion/Perspectives

- Muller avait raison.
- Kahan avait raison.

Conclusion/Perspectives

- Muller avait raison.
- Kahan avait raison.
- C'était plus compliqué que prévu (1 700 lignes).

Conclusion/Perspectives

- Muller avait raison.
- Kahan avait raison.
- C'était plus compliqué que prévu (1 700 lignes).
- On a des hypothèses précises pour lesquelles ça marche.

Conclusion/Perspectives

- Muller avait raison.
- Kahan avait raison.
- C'était plus compliqué que prévu (1 700 lignes).
- On a des hypothèses précises pour lesquelles ça marche.

- On a envie de tester/prouver les cas restants

Conclusion/Perspectives

- Muller avait raison.
- Kahan avait raison.
- C'était plus compliqué que prévu (1 700 lignes).
- On a des hypothèses précises pour lesquelles ça marche.

- On a envie de tester/prouver les cas restants
  - $p = 1$,

Conclusion/Perspectives

- Muller avait raison.
- Kahan avait raison.
- C'était plus compliqué que prévu (1 700 lignes).
- On a des hypothèses précises pour lesquelles ça marche.

- On a envie de tester/prouver les cas restants
  - $p = 1$,
  - $p = 2$,

Conclusion/Perspectives

- Muller avait raison.
- Kahan avait raison.
- C'était plus compliqué que prévu (1 700 lignes).
- On a des hypothèses précises pour lesquelles ça marche.

- On a envie de tester/prouver les cas restants
  - $p = 1$,
  - $p = 2$,
  - $p = 3$ et $\beta = 5$

Conclusion/Perspectives

- Muller avait raison.
- Kahan avait raison.
- C'était plus compliqué que prévu (1 700 lignes).
- On a des hypothèses précises pour lesquelles ça marche.

- On a envie de tester/prouver les cas restants
  - $p = 1$,
  - $p = 2$,
  - $p = 3$ et $\beta = 5$

  pour savoir s'il y a des contre-exemples ou si ce sont des artefacts de preuve.

Conclusion/Perspectives

- Muller avait raison.
- Kahan avait raison.
- C'était plus compliqué que prévu (1 700 lignes).
- On a des hypothèses précises pour lesquelles ça marche.

- On a envie de tester/prouver les cas restants
  - $p = 1$,
  - $p = 2$,
  - $p = 3$ et $\beta = 5$

  pour savoir s'il y a des contre-exemples ou si ce sont des artefacts de preuve.
- La gestion des arrondis au plus proche quelconques est assez pénible.

Conclusion/Perspectives

- Muller avait raison.
- Kahan avait raison.
- C'était plus compliqué que prévu (1 700 lignes).
- On a des hypothèses précises pour lesquelles ça marche.

- On a envie de tester/prouver les cas restants
  - $p = 1$,
  - $p = 2$,
  - $p = 3$ et $\beta = 5$

  pour savoir s'il y a des contre-exemples ou si ce sont des artefacts de preuve.
- La gestion des arrondis au plus proche quelconques est assez pénible.
- Underflow/Overflow : prouver un pre-scaling ?