Error in ulps of the multiplication or division by a correctly-rounded function or constant in binary floating-point arithmetic

N. Brisebarre J.-M. Muller J. Picot

CNRS and ENS de Lyon, Laboratoire LIP

Lyon, France

30th IEEE Symposium on Computer Arithmetic
Portland, OR, September 2023

Introduction

Goal: Tight (optimal or near optimal) error bounds in ulps for many usual functions:

$$\mathtt{x * pi},\quad \mathtt{ln(2)/x},\quad \mathtt{x/(y+z)},\quad \mathtt{(x+y)*z},\quad \mathtt{x/sqrt(y)},$$
$$\mathtt{sqrt(x)/y},\quad \mathtt{(x+y)(z+t)},\quad \mathtt{(x+y)/(z+t)},\quad \mathtt{(x+y)/(zt)},$$
$$\mathtt{(ax+b)/(cy+d)},\ \mathtt{(x*y)/sqrt(z)},\ \textit{etc.}$$

Context:

- radix-2, precision-p floating-point arithmetic, assuming round to nearest (any tie-breaking rule in the proofs, ties-to-even in the examples);

→ a FP number is zero or a number of the form $x = M_x \cdot 2^{e_x - p + 1}$, where $M_x, e_x \in \mathbb{Z}$, with $2^{p-1} \leqslant |M_x| \leqslant 2^p - 1$ (we assume no underflow or overflow);

- rounding function RN:

program line `z = x + y` $\Rightarrow$ obtained result $z = \mathrm{RN}(x + y)$.

Link between all these functions?

$\mathtt{x * pi}$, $\mathtt{ln(2)/x}$, $\mathtt{x/(y+z)}$, $\mathtt{(x+y)*z}$, $\mathtt{x/sqrt(y)}$,
$\mathtt{sqrt(x)/y}$, $\mathtt{(x+y)(z+t)}$, $\mathtt{(x+y)/(z+t)}$, $\mathtt{(x+y)/(zt)}$,
$\mathtt{(ax+b)/(cy+d)}$, $\mathtt{(x*y)/sqrt(z)}$, *etc.*

They are of the form

$$x \cdot c, \quad x/c, \quad c/x, \quad m \cdot n, \quad \text{or } n/d,$$

where

- x is a FPnumber, and
- c, n, m and d are either real constants or correctly-rounded functions of one or more variables.

Examples: $c = \pi$, or $c = \sqrt{y}$ where y is a FP number and $\sqrt{y}$ is obtained through the (correctly rounded) `sqrt` instruction, or $c = y + z$ obtained through `FPADD`.

Just an example

- program line

$$\texttt{t} = (\texttt{x} * \texttt{y})/\texttt{sqrt}(\texttt{z})$$

- real function

$$t = \frac{x \cdot y}{\sqrt{z}}$$

- computed result

$$\hat{t} = \mathrm{RN}\left(\frac{\mathrm{RN}(x \cdot y)}{\mathrm{RN}\left(\sqrt{z}\right)}\right)$$

We show that:

$$\left|t - \hat{t}\right| \leqslant \frac{5}{2}\,\mathrm{ulp}(t).$$

Very tight: $2.4994\,\mathrm{ulp}(t)$ attained in binary64 arithmetic.

Error in ulps vs. relative error

- numerical errors usually expressed as error in ulps or as relative errors.
- ulp(t) (*unit in the last place* of t) is $2^{\lfloor \log_2 |t| \rfloor - p + 1}$,
- if $t \neq 0$ is the exact result and $\hat{t}$ is the computed approximation:
  - the relative error is
  $$\left| \frac{t - \hat{t}}{t} \right|,$$
  - the error in ulps is
  $$\left| \frac{t - \hat{t}}{\mathrm{ulp}(t)} \right|.$$

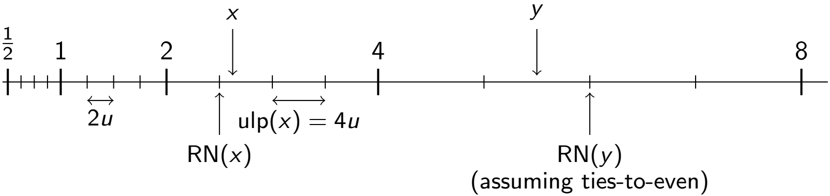
Error in ulps vs. relative error

- ulps preferred for "atomic" calculations (they convey more information: correct rounding almost equivalent to error $\leqslant 0.5\,\text{ulp}$);
- relative errors easier to manipulate for "large" calculations (e.g. from relative error on f and g, obtaining relative error on $f \times g$ is straightforward);
- easy conversion between both but at the cost of information loss:
  - define $u = 2^{-p}$ (unit roundoff);
  - we approximate an exact result t by a computed result $\hat{t}$:

$$
\begin{aligned}
&\text{error} \leqslant \alpha\,\text{ulp}(t) \\
&\qquad \Rightarrow \text{relative error} \leqslant 2\alpha u \\
&\qquad\qquad \Rightarrow \text{error} \leqslant 2\alpha\,\text{ulp}(t).
\end{aligned}
$$

$\rightarrow$ we have lost a factor 2 in the round trip conversion.

The FP numbers between 1/2 and 8 in the toy system $p = 3$



Multiplication of a FP number by a constant or a correctly-rounded function

Error bound in ulps on the computation of $x \cdot c$, where

- x is a FP number (assumed exact!) , and
- c is a real constant or a correctly-rounded function (can be $\sqrt{y}$, π, $y+z$, $y \cdot z$, etc.).

We want to bound the error of approximating $x \cdot c$ by

$$\mathrm{RN}\left(x \cdot \hat{c}\right),$$

where $\hat{c} = \mathrm{RN}(c)$. Here, we consider "general" bounds, applicable to any c.

[In the TETC paper we also try to improve these bounds in the particular case where c is a constant.]

Multiplication of a FP number by a constant or a correctly-rounded function

Property 1

Barring underflow and overflow, the FP number $s = \mathrm{RN}(\hat{c} \cdot x)$ *satisfies*

$$|s - cx| \leqslant \left(\frac{3}{2} - u\right) \cdot \mathrm{ulp}\,(cx) < \frac{3}{2}\,\mathrm{ulp}\,(cx). \qquad \square$$

In the general case (arbitrary constant c) the bound is asymptotically optimal. Shown with the following generic example (assuming RN breaks ties to even):

If p is even, choose

$$\begin{aligned} x &= 2^p - 2^{p/2}, \\ c &= 1 + 2^{-p/2-1} - 2^{-p}, \end{aligned}$$

If p is odd, choose

$$\begin{aligned} x &= 2^p - 2^{(p-1)/2}, \\ c &= 1 + 2^{-(p+1)/2} - 2^{-p}. \end{aligned}$$

Some examples

- previous example: c can be expressed as sum of two FPNs
  $\rightarrow$ asymptotic optimality of the bound 1.5 ulp of Property 1 for the calculation of $\mathtt{z} * (\mathtt{x} + \mathtt{y})$;

- errors

  $$\begin{array}{ll} 1.499756 \cdots \mathrm{ulp}(cx) & \text{(binary32 arithmetic),} \\ 1.499999992549 \cdots \mathrm{ulp}(cx) & \text{(binary64 arithmetic)} \\ 1.49999999999999999993061 \cdots \mathrm{ulp}(cx) & \text{(binary128 arithmetic)} \end{array}$$

  can be attained when calculating $\mathtt{z} * (\mathtt{x} * \mathtt{y})$, showing that for that function, the bound is very tight;

- in binary64 arithmetic, with $x = 9007197761440759$ and $y = 4503599630388691/2^{52}$ error when computing $x\sqrt{y}$ is $1.4991\,\mathrm{ulp}(x\sqrt{y})$.

Division of a FP number by a correctly-rounded function

We approximate x/c, where x is aFP number and c is either a real constant or a real function of one or more FP variables, by

$$s = \mathrm{RN}(x/\hat{c}),$$

where, as previously, $\hat{c} = \mathrm{RN}(c)$.

Property 2

Barring underflow and overflow, the FP number $s = \mathrm{RN}(x/\hat{c})$ *satisfies*

$$\begin{aligned}\left|s - \frac{x}{c}\right| &\leqslant \left(\frac{3}{2} - \frac{2u}{1+2u}\right) \mathrm{ulp}\left(\frac{x}{c}\right) \\ &\leqslant \left(\frac{3}{2} - 2u + 4u^2\right) \mathrm{ulp}\left(\frac{x}{c}\right) \\ &< \frac{3}{2} \mathrm{ulp}\left(\frac{x}{c}\right).\end{aligned}$$

□

As for the product, "generic" example for a general constant c that shows asymptotic optimality.

Tightness?

- asymptotic optimality of the bound $1.5\,\mathrm{ulp}$ for the calculation of $\mathtt{z}/(\mathtt{x}+\mathtt{y})$.
- errors
  $$\left.\begin{array}{ll} 1.49957\cdots\mathrm{ulp}(x/c) & \text{(binary32)},\\ 1.49999998137\cdots\mathrm{ulp}(x/c) & \text{(binary64)},\\ 1.49999999999999998265\cdots\mathrm{ulp}(x/c) & \text{(binary128)}\end{array}\right.\quad \text{can be}$$
  attained when calculating $\mathtt{z}/(\mathtt{x}*\mathtt{y})$, showing that for that function, the bound is very tight;
- binary64 arithmetic, error $1.49906\,\mathrm{ulp}(x/\sqrt{y})$ attained for $x = 9007198105271337$ and $y = 4503599631275935/2^{52}$ when calculating $x/\sqrt{y}$.

Dividing a correctly-rounded function by a FP number

Now we consider approximating c/x, where x is a FP number and c is either a real constant or a real function of one or more FP variables, by

$$s = \mathrm{RN}(\hat{c}/x),$$

where, as previously, $\hat{c} = \mathrm{RN}(c)$.

Property 3

Barring underflow and overflow, the FP number $s = \mathrm{RN}(\hat{c}/x)$ *satisfies*

$$\begin{aligned} \left| s - \frac{c}{x} \right| &\leqslant \frac{3 + 2u}{2 + 4u} \cdot \mathrm{ulp}\left(\frac{c}{x}\right) \\ &\leqslant \left(\frac{3}{2} - 2u + 4u^2\right) \mathrm{ulp}\left(\frac{c}{x}\right). \end{aligned}$$

□

Similar examples of asymptotic optimality or tightness. Covers functions such as $\ln(2)/x$, $\sqrt{x}/y$, $(x + y)/z$, ...

Product of two correctly-rounded functions

Approximation of $m \cdot n$, where m and n are either real constants or correctly-rounded functions, by

$$s = \mathrm{RN}\left(\hat{m} \cdot \hat{n}\right),$$

where $\hat{m} = \mathrm{RN}(m)$ and $\hat{n} = \mathrm{RN}(n)$ (of course nobody multiplies 2 constants)

Property 4

Barring underflow and overflow, the FP number $s = \mathrm{RN}(\hat{m} \cdot \hat{m})$ *satisfies*

$$|s - mn| \leqslant \left(\frac{5}{2} + \frac{u}{2}\right) \mathrm{ulp}\left(mn\right). \qquad \square$$

In the general case, the bound is asymptotically optimal for even values of p (it probably is for odd values too but no proof).

Tightness and examples of application

- error $2.4999982\,\mathrm{ulp}(efgh)$ is attained when computing $(\mathtt{e} * \mathtt{f}) * (\mathtt{g} * \mathtt{h})$ in binary64/double-precision arithmetic,
- the property applies to calculations such as $\pi \cdot \sqrt{x}$, $(x+y)\cdot(z+t)$, $(x \cdot y)\cdot\sqrt{z}$, $e^x \cos(y)$ (with correctly rounded functions), etc. If an FMA instruction is available, it also covers computations of the form

$$(ax+b)(cy+d),$$

where a, b, c, d, x, and y are FP numbers.

Quotient of two correctly-rounded functions

Approximation of n/d, where n and d are either real constants or correctly-rounded functions, by

$$s = \mathrm{RN}\left(\frac{\hat{n}}{\hat{d}}\right),$$

where $\hat{n} = \mathrm{RN}(n)$ and $\hat{d} = \mathrm{RN}(d)$.

Property 5

Barring underflow and overflow, the floating-point number $s = \mathrm{RN}(\hat{n}/\hat{d})$ *satisfies*

$$\left|s - \frac{n}{d}\right| \leqslant \frac{5}{2}\,\mathrm{ulp}\left(\frac{n}{d}\right). \qquad \square$$

covers calculations such as $\pi/\sqrt{x}$, $(x+y)/(z+t)$, $(xy)/(z+t)$, etc. If an FMA instruction is available, it also covers computations of the form

$$\frac{ax+b}{cy+d},$$

where a, b, c, d, x, and y are FP numbers.

Tightness?

- binary64, error $2.49999997392\cdots$ ulp attained when computing $(x+y)/(z+t)$, $(xy)/(z+t)$; $(x+y)/(zt)$, and $(xy)/(zt)$ with well chosen values (see TETC paper);
- binary64, error 2.4994 ulp attained when computing

$$\frac{x+y}{\sqrt{z}}$$

or

$$\frac{xy}{\sqrt{z}},$$

with well chosen values.

Conclusion

- sharp error bounds in ulps for computations in binary FP arithmetic of the form $x \cdot c$, x/c, c/x, $m \cdot n$ and n/d, where x is a FP number and c, n, m and d are either real constants or correctly-rounded functions of one or more variables;
- examples of functions for which our work gives tight bounds are

$$\mathtt{x*pi},\ \mathtt{ln(2)/x},\ \mathtt{x/(y+z)},\ \mathtt{(x+y)*z},\ \mathtt{x/sqrt(y)},\ \mathtt{sqrt(x)/y},\\ \mathtt{(x+y)*(z+t)},\ \mathtt{(x+y)/(z+t)},\ \mathtt{(x+y)/(zt)},\\ \mathtt{(ax+b)/(cy+d)},\ \mathtt{(x*y)*sqrt(z)},\ \text{etc.}$$

- In several cases, we have been able to show that our bounds are asymptotically optimal.

Thank you!