# Supplementary material to "Accelerating Correctly Rounded Floating-Point Division When the Divisor is Known in Advance"

Nicolas Brisebarre, Jean-Michel Muller and Saurabh Kumar Raina
Laboratoire LIP, ENSL/CNRS/INRIA Arenaire Project
Ecole Normale Supérieure de Lyon
46 Allée d'Italie, 69364 Lyon Cedex 07, FRANCE

Nicolas.Brisebarre@ens-lyon.fr, Jean-Michel.Muller@ens-lyon.fr, Saurabh-Kumar.Raina@ens-lyon.fr

December 10, 2003

## Appendix: tables, proofs and intermediate lemmas

We will frequently use the two following well-known properties, whose proofs are straightforward:

**Property 5**

- *Let $y \in \mathbb{M}_n$. There exists $q \in \mathbb{N}$ such that $1/y$ belongs to $\mathbb{M}_q$ if and only if $y$ is a power of 2.*
- *If $m > n$, the exact quotient of two $n$-bit numbers cannot be an $m$-bit number.*
- *Let $x, y \in \mathbb{M}_n$. $x \neq y \Rightarrow |x/y - 1| \geq 2^{-n}$.*

We call a **breakpoint** a value $z$ where the rounding changes, that is, if $t_1$ and $t_2$ are real numbers satisfying $t_1 < z < t_2$ and $\circ_t$ is the rounding mode, then $\circ_t(t_1) < \circ_t(t_2)$. For "directed" rounding modes (i.e., towards $+\infty$, $-\infty$ or $0$), the breakpoints are the FP numbers. For rounding to the nearest mode, they are the exact middle of two consecutive FP numbers.

For $a \in \mathbb{M}_n$, we define $a^+$ as its **successor** in $\mathbb{M}_n$, that is, $a^+ = \min\{b \in \mathbb{M}_n, b > a\}$, and $a^-$ as the **predecessor of** $a$, that is, $a^- = \max\{b \in \mathbb{M}_n, b < a\}$.

The next result gives a lower bound on the distance between a breakpoint (in round-to-nearest mode) and the quotient of two FP numbers.

**Property 6** *If $x, y \in \mathbb{M}_n$, $1 \leq x, y < 2$, then the distance between $x/y$ and the middle of two consecutive FP numbers is lower-bounded by $\frac{1}{y \times 2^{2n-1}} > \frac{1}{2^{2n}}$ if $x \geq y$; and $\frac{1}{y \times 2^{2n}} > \frac{1}{2^{2n+1}}$ otherwise. Moreover, if the last mantissa bit of $y$ is a zero, then the lower bounds become twice these ones.*

**Proof of Property 1.** Let $x, y \in \mathbb{M}_n$. Without loss of generality, we can assume that $x$ and $y$ belong to $[1, 2)$. Since the cases $x, y = 1$ or 2 are straightforward, we assume that $x$ and $y$ belong to $(1, 2)$. Thus $1/y \notin \mathbb{M}_n$. Since $z_h = \circ_\nu(z)$ and $z \in (1/2, 1)$, we have, $\left|\frac{1}{y} - z_h\right| < 2^{-n-1}$. Therefore,

$$\left|\frac{x}{y} - xz_h\right| < 2^{-n}. \tag{1}$$

From Property 5 and (1), we cannot have $x/y > 1$ and $xz_h < 1$ or the converse. So $xz$ and $xz_h$ belong to the same "binade" (i.e., $\text{ulp}(xz_h) = \text{ulp}(xz)$). Now, there are two possible cases:

- if $x \geq y$, then $|xz_h - \circ_\nu(xz_h)| \leq 2^{-n}$, so $|x/y - \circ_\nu(xz_h)| < 2^{-n+1} = \text{ulp}(x/y)$.
- if $x < y$, then $|xz_h - \circ_\nu(xz_h)| \leq 2^{-n-1}$, so $|x/y - \circ_\nu(xz_h)| < 3 \times 2^{-n-1} = 1.5 \times \text{ulp}(x/y)$.

□

To analyze the behavior of Algorithm 1, we will need the following property.

**Property 7** *If $x < y$ and $1 \leq x, y < 2$, then the naive solution returns a result $q$ such that either $q$ is within 1 ulp from $x/y$, or $x/y$ is at least at a distance $\frac{2^{-2n+1}}{y} + 2^{-2n+1} - \frac{2^{-3n+2}}{y}$ from a breakpoint of the round-to-nearest mode.*

**Proof of Property 7.** The proof is similar to that of Property 1. We use the tighter bounds:

- $|1/y - z_h| < 2^{-n-1} - 2^{-2n}/y$ (this comes from Property 6: $1/y$ is at a distance at least $2^{-2n}/y$ from a breakpoint);
- $x \leq 2 - 2^{-n+2}$ (this comes from $x < y < 2$, which implies $x \leq (2^-)^-$).

Combining these bounds gives

$$\left|\frac{x}{y} - xz_h\right| \leq 2^{-n} - \frac{2^{-2n+1}}{y} - 2^{-2n+1} + \frac{2^{-3n+2}}{y}.$$

The final bound $\ell_{min}$ is obtained by adding the $1/2$ ulp bound on $|xz_h - \circ_\nu(xz_h)|$:

$$\left|\frac{x}{y} - \circ_\nu(xz_h)\right| \leq \ell_{min} = 3 \times 2^{-n-1} - \frac{2^{-2n+1}}{y} - 2^{-2n+1} + \frac{2^{-3n+2}}{y}.$$

If $\circ_\nu(xz_h)$ is not within 1 ulp from $x/y$, it means that $x/y$ is at a distance at least $1/2$ ulp from the breakpoints that are immediately above or below $q = \circ_\nu(xz_h)$. And since the breakpoints that are immediately above $\circ_\nu(xz_h)^+$ or below $\circ_\nu(xz_h)^-$ are at a distance $1.5$ ulps $= 3 \times 2^{-n-1}$ from $\circ_\nu(xz_h)$, $x/y$ is at least at a distance $3 \times 2^{-n-1} - \ell_{min}$ from these breakpoints. □

**Proof of Property 2**. We look for the couples $(x, y) \in \mathbb{M}_n$ such that $1 \leq x < y < 2$ and

$|x/y - \circ_\nu (x \circ_\nu (1/y))|$ is as close as possible to $\frac{1.5}{2^n}$. To hasten the search, we will look for couples such that $\left|\frac{x}{y} - \circ_\nu \left(x \circ_\nu \left(\frac{1}{y}\right)\right)\right| \geq \frac{2K+1}{2^{n+1}}$, where $K$ is a real parameter as close as possible to 1. If we write

$$\frac{x}{y} - \circ_\nu \left(x \circ_\nu \left(\frac{1}{y}\right)\right) = \frac{x}{y} - x \circ_\nu \left(\frac{1}{y}\right) + x \circ_\nu \left(\frac{1}{y}\right) - \circ_\nu \left(x \circ_\nu \left(\frac{1}{y}\right)\right),$$

we see that, as $\left|x \circ_\nu \left(\frac{1}{y}\right) - \circ_\nu \left(x \circ_\nu \left(\frac{1}{y}\right)\right)\right| \leq \frac{1}{2}\frac{1}{2^n}$, we want

$$x \left|\frac{1}{y} - \circ_\nu \left(\frac{1}{y}\right)\right| \geq \frac{K}{2^n} \tag{2}$$

Hence $x > 2K$ since $\left|\frac{1}{y} - \circ_\nu \left(\frac{1}{y}\right)\right| < \frac{1}{2^{n+1}}$ ($1/y \notin \mathbb{M}_n$). Let us write $y = \frac{2^n - s}{2^{n-1}}$ and $x = \frac{2^n - l}{2^{n-1}}$ with $1 \leq s \leq \lfloor 2^n(1-K) \rfloor - 1$ and $s+1 \leq l \leq \lfloor 2^n(1-K) \rfloor$. We have,

$$\frac{2^n}{y} = 2^{n-1} + \frac{s}{2} + \frac{1}{2}\frac{s^2}{2^n - s}.$$

As $y > x$, (2) implies

$$\left|\frac{1}{y} - \circ_\nu \left(\frac{1}{y}\right)\right| > \frac{K}{2^n y}. \tag{3}$$

The full proof considers two cases: $s$ is odd and $s$ is even. For reasons of space we only deal with the case "$s$ odd" here. The other case is very similar (the full proof can be obtained through an email to one of the authors).

When $s$ is odd, we only keep the $s \in [1, \lfloor 2^n(1-K) \rfloor - 1]$ such that

$$\frac{1}{2}\frac{s^2}{2^n - s} \in \left(0, \frac{1}{2} - K\frac{2^{n-1}}{2^n - s}\right) \cup \bigcup_{k \in \mathbb{N}\setminus\{0\}} \left(k + K\frac{2^{n-1}}{2^n - s} - \frac{1}{2}, k + \frac{1}{2} - K\frac{2^{n-1}}{2^n - s}\right)$$

i.e., $s \in [1, \lfloor 2^n(1-K) \rfloor - 1] \cap \bigcup_{k \in \mathbb{N}} (a_{\text{odd},k}, b_{\text{odd},k})$ with

$$a_{\text{odd},0} = 0 \text{ and } a_{\text{odd},k} = \frac{-2k + 1 + \sqrt{(2k-1)^2 + 2^{n+2}(2k - 1 + K)}}{2} \text{ for all } k \geq 1,$$

$$b_{\text{odd},k} = \frac{-2k - 1 + \sqrt{(2k+1)^2 + 2^{n+2}(2k + 1 - K)}}{2} \text{ for all } k.$$

Let $k_{\text{odd}} = \max\{k \in \mathbb{N}, a_{\text{odd},k} < \lfloor 2^n(1-K) \rfloor - 1\}$. We have

$$k_{\text{odd}} = \left\lfloor \frac{1}{2}\frac{2^{n+2}(1-K) + 4(\lfloor 2^n(1-K) \rfloor - 1)(\lfloor 2^n(1-K) \rfloor - 2)}{2^{n+2} - 4\lfloor 2^n(1-K) \rfloor + 4} \right\rfloor.$$

Finally, when $s$ is odd, we only keep the

$$s \in \bigcup_{0 \leq k \leq k_{\text{odd}} - 1} (a_{\text{odd},k}, b_{\text{odd},k}) \cup (a_{\text{odd},k_{\text{odd}}}, \min(b_{\text{odd},k_{\text{odd}}}, \lfloor 2^n(1-K) \rfloor)).$$

Let $k \in \mathbb{N}$, $0 \leq k \leq k_{\text{odd}}$ such that $s \in (a_{\text{odd},k}, b_{\text{odd},k})$. We have $2^n \circ_\nu \left(\frac{1}{y}\right) = 2^{n-1} + \frac{s \pm 1}{2} + k$, with $\pm = +$ if $s > -k + \sqrt{k^2 + 2^{n+1}k}$ and $\pm = -$ otherwise. Thus,

$$2^n x \circ_\nu \left(\frac{1}{y}\right) = 2^n - l + s \pm 1 + 2k - \frac{l(s \pm 1 + 2k)}{2^n}. \tag{4}$$

Now, recall that we want

$$\left| x \circ_\nu \left(\frac{1}{y}\right) - \circ_\nu \left(x \circ_\nu \left(\frac{1}{y}\right)\right)\right| \geq \frac{2K+1}{2^{n+1}} - \left|\frac{x}{y} - x \circ_\nu \left(\frac{1}{y}\right)\right|. \tag{5}$$

This can be written as

$$\left| x \circ_\nu \left(\frac{1}{y}\right) - \circ_\nu \left(x \circ_\nu \left(\frac{1}{y}\right)\right)\right| \geq \frac{2K+1}{2^{n+1}} - \frac{2^n - l}{2^{2n-1}} \left|\frac{1}{2}\frac{s^2}{2^n - s} - \frac{(2k \pm 1)}{2}\right| = \varepsilon_{s,l,k,K}.$$

We get from this condition and (4), that

$$\frac{l(s \pm 1 + 2k)}{2^n} \in \bigcup_{m \in \mathbb{N}} (m + 2^n \varepsilon_{s,l,k,K}, m + 1 - 2^n \varepsilon_{s,l,k,K},)$$

i.e., $l \in [s+1, \lfloor 2^n(1-K) \rfloor] \cap \bigcup_{m \in \mathbb{N}} (c_{\text{odd},m}, d_{\text{odd},m})$ where

$$c_{\text{odd},m} = \frac{2^n(m + K + 1/2) - 2^n |s^2/(2^n - s) - (2k \pm 1)|}{s \pm 1 + 2k - |s^2/(2^n - s) - (2k \pm 1)|}$$

and $d_{\text{odd},m} = \dfrac{2^n(m - K + 1/2) + 2^n |s^2/(2^n - s) - (2k \pm 1)|}{s \pm 1 + 2k + |s^2/(2^n - s) - (2k \pm 1)|}$.

Let $m_{\text{odd}} = \min\{m \in \mathbb{N}, s < d_{\text{odd},m}\}$ and $M_{\text{odd}} = \max\{m \in \mathbb{N}, c_{\text{odd},m} < \lfloor 2^n(1-K) \rfloor\}$. We easily get an exact expression of these integers. Hence, we look for the

$$l \in (\max(c_{\text{odd},m_{\text{odd}}}, s), d_{\text{odd},m_{\text{odd}}}) \cup \bigcup_{m_{\text{odd}}+1 \leq m \leq M_{\text{odd}}-1} (c_{\text{odd},m}, d_{\text{odd},m}) \cup (c_{\text{odd},M_{\text{odd}}}, \min(d_{\text{odd},M_{\text{odd}}}, \lfloor 2^n(1-K) \rfloor)).$$

Once we have got all these couples $(s,l)$, we end up our research by checking if $\left|\frac{x}{y} - \circ_\nu \left(x \circ_\nu \left(\frac{1}{y}\right)\right)\right| \geq \frac{2K+1}{2^{n+1}}$ with $x = (2^n - l)/2^{n-1}$ and $y = (2^n - s)/2^{n-1}$.

These remarks lead to an algorithm implemented in GP, the calculator of PARI [5], that gets faster as the parameter $K$ grows. If $K$ is too large, we won't find any couple. But, we know values of $K$ that are close to 1 and associated to a couple $(x,y)$. These values allow us to get the couples $(x,y) \in \mathbb{M}_n$ such that $1 \leq x < y < 2$ and $\left|\frac{x}{y} - \circ_\nu \left(x \circ_\nu \left(\frac{1}{y}\right)\right)\right|$ is as close as possible to $\frac{1.5}{2^n}$. More precisely, we now give a

sequence $(x_n, y_n)_{n\in\mathbb{N}\setminus\{0\}}$ such that, for all $n \in \mathbb{N}\setminus\{0\}$, $x_n,\ y_n \in \mathbb{M}_n$, $1 \le x_n < y_n < 2$ and $2^n \left| \dfrac{x_n}{y_n} - \circ_\nu \left( x_n \circ_\nu \left( \dfrac{1}{y_n} \right) \right) \right| \longrightarrow \dfrac{3}{2}$ as $n \longrightarrow +\infty$. For $n$ even, we choose

$$x_n = \frac{2^n - 2^{n/2} - 2^{n/2-1} + 3}{2^{n-1}}, \quad y_n = \frac{2^{n/2} - 1}{2^{n/2-1}}.$$

For $n$ odd, we choose

$$x_n = \frac{2^{(n+3)/2} - 7}{2^{(n+1)/2}}, \quad y_n = \frac{2^n - 2^{(n+1)/2} + 1}{2^{n-1}}.$$

Let $n = 2p$, $p \in \mathbb{N}\setminus\{0\}$. We have $\dfrac{x_{2p}}{y_{2p}} = \frac{2^{p-1}}{2^p-1} \dfrac{2^{2p} - 2^p - 2^{p-1} + 3}{2^{2p-1}}$. After some calculation, we get, for all $p \ge 2$,

$$2^{2p} \left| \frac{x_{2p}}{y_{2p}} - \circ_\nu \left( x_{2p} \circ_\nu \left( \frac{1}{y_{2p}} \right) \right) \right| = \left| \frac{3}{2} - \frac{5}{2} \frac{2^{-p}}{1 - 2^{-p}} \right| \longrightarrow \frac{3}{2} \text{ as } p \longrightarrow +\infty.$$

Let $n = 2p + 1$, $p \in \mathbb{N}$. We have $\dfrac{x_{2p+1}}{y_{2p+1}} = \frac{2^{p+2}-7}{2^{p+1}} \frac{2^{2p}}{2^{2p+1}-2^{p+1}+1}$. After some calculation, we get, for all $p \ge 2$,

$$\begin{aligned} &2^{2p+1} \left| \frac{x_{2p+1}}{y_{2p+1}} - \circ_\nu \left( x_{2p+1} \circ_\nu \left( \frac{1}{y_{2p+1}} \right) \right) \right| \\ &= \left| \frac{3}{2} - 7 \cdot 2^{-p-2} - (2^{-2p-1} - 7 \cdot 2^{-3p-3}) \frac{1}{1 - 2^{-p} + 2^{-2p-1}} \right| \longrightarrow \frac{3}{2} \text{ as } p \longrightarrow +\infty. \end{aligned}$$

Then we use our algorithm with the parameter $K$ obtained from this sequence. We get the values given in Table 1. Note that the couples $(x, y)$ in the table are the couples $(x_n, y_n)$ except for $n = 64$. □

**Sketch of a proof for Conjecture 1.** Define $z = 1/y = z_h + z_\rho$, where $z_h = \circ_\nu(z)$, with $1 < y < 2$. When $n \to \infty$, the maximum value of $|z_\rho|$ is asymptotically equal to $1/2\text{ulp}(z)$, and its average value is asymptotically equal to $1/4\text{ulp}(z) = 2^{-n-2}$. Hence, for $1 < x < 2$, we can write: $xz = xz_h + \epsilon$ where the average value of $|\epsilon|$ is $\dfrac{y+1}{2} \times 2^{-n-2} = (y+1)2^{-n-3}$ for $x < y$ and $\dfrac{2+y}{2} \times 2^{-n-2} = (2+y)2^{-n-3}$ for $x > y$ (to get these figures, we multiply the average value of $\epsilon$ by the average value of $x$, which is $\frac{y+1}{2}$ for $1 < x < y$ and $\frac{2+y}{2}$ for $y < x < 2$). The "breakpoints" of the rounding mode[1], are regularly spaced, at distance $2^{-n}$ for $x < y$, and $2^{-n+1}$ for $x > y$. Therefore, the probability that $\circ_\nu(xz) \neq \circ_\nu(xz_h)$ should asymptotically be the probability that there should be a breakpoint between these values. That probability is $(y+1)2^{-n-3}/2^{-n} = \dfrac{y+1}{8}$ for $x < y$, and $(2+y)2^{-n-3}/2^{-n+1} = \dfrac{y+2}{16}$ for $x > y$.

[1]Since we assume rounding to nearest mode, the breakpoints are the exact middles of two consecutive machine numbers.

Therefore, for a given $y$, the probability that the naive method should give a result different from $\circ_\nu(x/y)$ is $\frac{(y+1)(y-1)}{8} + \frac{(y+2)(2-y)}{16} = \frac{y^2}{16} + \frac{1}{8}$. Therefore, assuming now that $y$ is variable, the probability that the naive method give an incorrectly rounded result is

$$\int_1^2 \left(\frac{y^2}{16} + \frac{1}{8}\right) dy = \frac{13}{48} \approx 0.27.$$

□

The following result, due to Markstein, was designed in order to get a correctly rounded result from an approximation to a quotient obtained using Newton-Raphson or Goldschmidt iterations. We give it here, since we use it in the proof of Theorem 1.

**Theorem 4 (Markstein, 1990 [3, 4])** *Assume $x, y \in \mathbb{M}_n$. If $u \in \mathbb{M}_n$ is within $1/2$ ulp of $1/y$ and $q \in \mathbb{M}_n$, $q$ within $1$ ulp of $x/y$ then one application of*

$$\begin{cases} r &= \circ_\nu(x - qy) \\ q' &= \circ_\nu(q + ru) \end{cases}$$

*yields $q' = \circ_\nu(x/y)$.*

One would like to use Theorem 4 to get a correctly rounded result from an initial value $q$ obtained by the naive method, that is, by computing $\circ_\nu(xz_h)$, where $z_h = \circ_\nu(1/y)$. Unfortunately, $q$ will not always be within one ulp from $x/y$ (see Property 1), so Theorem 4 cannot be directly applied. One could get a better initial approximation to $x/y$ by performing one step of Newton-Raphson iteration from $q$. And yet, such an iteration step is not necessary, as shown by Theorem 1.

**Proof of Theorem 1.** We assume $1 \leq x, y < 2$. First, let us notice that if $x \geq y$, then (from Property 1), $q$ is within one ulp from $x/y$, therefore Theorem 4 applies, hence $q' = \circ_\nu(x/y)$. Let us now focus on the case $x < y$. Define $\epsilon_1 = x/y - q$ and $\epsilon_2 = 1/y - z_h$. From Property 1 and the definition of rounding to nearest, we have, $|\epsilon_1| < 3 \times 2^{-n-1}$ and $|\epsilon_2| < 2^{-n-1}$. The number $\rho = x - qy = \epsilon_1 y$ is less than $3 \times 2^{-n}$ and is a multiple of $2^{-2n+1}$. It therefore can be represented exactly with $n+1$ bits of mantissa. Hence, the difference between that number and $r = \circ_\nu(x - qy)$ (i.e., $\rho$ rounded to $n$ bits of mantissa) is zero or $\pm 2^{-2n+1}$. Therefore, $r = \epsilon_1 y + \epsilon_3,$ with $\epsilon_3 \in \{0, \pm 2^{-2n+1}\}$.

Let us now compute $q + rz_h$. We have $q + rz_h = \dfrac{x}{y} + \dfrac{\epsilon_3}{y} - \epsilon_1\epsilon_2 y - \epsilon_2\epsilon_3$. Hence,

$$\left|\frac{x}{y} - (q + rz_h)\right| \leq \frac{2^{-2n+1}}{y} + 3 \times 2^{-2n-2} y + 2^{-3n}$$

Define $\epsilon = 2^{-2n+1}/y + 3 \times 2^{-2n-2} y + 2^{-3n}$. Now, from Property 7, either $q$ was at a distance less than one ulp from $x/y$ (but in such a case, $q' = \circ_\nu(x/y)$ from Theorem 4), or $x/y$ is at least at a distance

$$\delta = \frac{2^{-2n+1}}{y} + 2^{-2n+1} - \frac{2^{-3n+2}}{y}.$$

from a breakpoint. A straightforward calculation shows that, if $n \geq 4$, then $\epsilon < \delta$. Therefore there is no breakpoint between $x/y$ and $q+rz_h$. Hence $\circ_\nu(q+rz_h) = \circ_\nu(x/y)$. The cases $n < 4$ are easily checked through exhaustive testing. □

**Proof of Property 3.** Without any loss of generality, we assume $1 \leq x, y < 2$. Define $K = n + 1$ if $q < 1$ and $K = n$ otherwise. Since $r$ is a multiple of $2^{-n-K+2}$ that is less than $2^{-K+1}y$, we have $r \in \mathbb{M}_n$. Hence, it is computed exactly. □

**Proof of Property 4.** From Property 3, $1 - yz_h$ is computed exactly. Therefore $\rho$ is *exactly* equal to $1-yz_h$. Hence, $\rho/y$ is equal to $1/y-z_h$. Hence, $z_\ell$ is equal to $\circ_\nu(1/y-z_h)$. □

**Proof of Theorem 2.** Without loss of generality, we assume $x \in (1, 2)$ and $y \in (1, 2)$ (the cases $x = 1$ or $y = 1$ are straightforward). This gives $z \in (1/2, 1)$, and, from Property 5, the binary representation of $z$ is infinite. Hence, $z_h \in [1/2, 1]$. The case $z_h = 1$ is impossible ($y > 1$ and $y \in \mathbb{M}_n$ imply $y \geq 1 + 2^{-n+1}$, thus $1/y \leq 1 - 2^{-n+1} + 2^{-2n+2} < 1 - 2^{-n} \in \mathbb{M}_n$, thus $\circ_\nu(1/y) \leq 1 - 2^{-n}$). Hence, the binary representation of $z_h$ has the form $0.z_h^1 z_h^2 z_h^3 \cdots z_h^n$. Since $z_h$ is obtained by rounding $z$ to the nearest, we have: $|z - z_h| \leq \frac{1}{2}\text{ulp}(z) = 2^{-n-1}$. Moreover, Property 5 shows that the case $|z - z_h| = 2^{-n-1}$ is impossible. Therefore $|z - z_h| < 2^{-n-1}$. From this, we deduce: $|z_\ell| = |\circ_\nu(z - z_h)| \leq 2^{-n-1}$. Again, the case $|z_\ell| = 2^{-n-1}$ is impossible: if we had $|z - z_h| < 2^{-n-1} = |z_\ell|$, this would imply $|z - (z_h + 2^{-n-1})| < 2^{-2n-1}$ or $|z - (z_h - 2^{-n-1})| < 2^{-2n-1}$ which would contradict the fact that the binary representation of the reciprocal of an $n$-bit number cannot contain more than $n - 1$ consecutive zeros or ones [1, 2]. Therefore $|z_\ell| < 2^{-n-1}$. Thus, from the definition of $z_\ell$, $|(z - z_h) - z_\ell| < 2^{-2n-2}$. This implies $|x(z - z_h) - xz_\ell| < 2^{-2n-1}$, hence, $|x(z - z_h) - \circ_\nu(xz_\ell)| < 2^{-2n-1} + \frac{1}{2}\text{ulp}(xz_\ell) \leq 2^{-2n}$. Therefore,

$$|xz - \circ_\nu[xz_h + \circ_\nu(xz_\ell)]| < 2^{-2n} + \frac{1}{2}\text{ulp}\left(xz_h + \circ_\nu(xz_\ell)\right) \tag{6}$$

Hence, if for a given $y$ there does not exist any $x$ such that $x/y = xz$ is at a distance less than $2^{-2n}$ from the middle of two consecutive FP numbers, then $\circ_\nu[xz_h + \circ_\nu(xz_\ell)]$ will always be equal to $\circ_\nu(xz)$, i.e., Algorithm 2 will give a correct result. Therefore, from Property 6, if $x \geq y$ then Algorithm 2 will return a correctly rounded quotient. Also, if $|z_\ell| < 2^{-n-2}$ (which corresponds to Condition "$|z_\ell| < 2^{-n-e-2}$" of the theorem we are proving) then we get a sharper bound:

$$|xz - \circ_\nu[xz_h + \circ_\nu(xz_\ell)]| < 2^{-2n-1} + \frac{1}{2}\text{ulp}\left(xz_h + \circ_\nu(xz_\ell)\right) \tag{7}$$

and Property 6 implies that we get a correctly rounded quotient.

Let us now focus on the case $x < y$. Let $q \in \mathbb{M}_n$, $1/2 \leq q < 1$, and define integers $X$, $Y$ and $Q$ as

$$\left\{ \begin{array}{rcl} X & = & x \times 2^{n-1}, \\ Y & = & y \times 2^{n-1}, \\ Q & = & q \times 2^{n}. \end{array} \right.$$

If we have $\frac{x}{y} = q + 2^{-n-1} + \epsilon,$ with $|\epsilon| < 2^{-2n}$, then

$$2^{n+1}X = 2QY + Y + 2^{n+1}\epsilon Y, \text{ with } |\epsilon| < 2^{-2n}. \tag{8}$$

But:

- Equation (8) implies that $R' = 2^{n+1}\epsilon Y$ should be an integer.
- The bounds $Y < 2^n$ and $|\epsilon| < 2^{-2n}$ imply $|R'| < 2$.
- Property 5 implies $R' \neq 0$.

Hence, the only possibility is $R' = \pm 1$. Therefore, to find values $y$ for which for any $x$ Algorithm 2 gives a correct result, we have to examine the possible integer solutions to

$$\begin{cases} 2^{n+1}X = (2Q+1)Y \pm 1, \\ 2^{n-1} \leq X \leq 2^n - 1, \\ 2^{n-1} \leq Y \leq 2^n - 1, \\ 2^{n-1} \leq Q \leq 2^n - 1. \end{cases} \tag{9}$$

There are no solutions to (9) for which $Y$ is even. This shows that if the last mantissa bit of $y$ is a zero, then Algorithm 2 always returns a correctly rounded result. Now, if $Y$ is odd then it has a reciprocal modulo $2^{n+1}$. Define $P_- = (1/Y) \bmod 2^{n+1}$ and $P_+ = (-1/Y) \bmod 2^{n+1}$, $Q_- = (P_- - 1)/2$ and $Q_+ = (P_+ - 1)/2$. From $0 < P_-, P_+ \leq 2^{n+1} - 1$ and $P_- + P_+ = 0 \bmod 2^{n+1}$, we easily find $P_- + P_+ = 2^{n+1}$. From this, we deduce,

$$\begin{array}{l} 0 \leq Q_-, Q_+ \leq 2^n - 1, \\ Q_- + Q_+ = 2^n - 1. \end{array} \tag{10}$$

Define $X_- = \frac{P_- \times Y - 1}{2^{n+1}}$ and $X_+ = \frac{P_+ \times Y + 1}{2^{n+1}}$. From (10) we easily deduce that either $Q_- \geq 2^{n-1}$ or $Q_+ \geq 2^{n-1}$, but both are impossible. Hence, either $(Y, X_+, Q_+)$ or $(Y, X_-, Q_-)$ can be solution to Eq. (9), but both are impossible. Algorithm 3 checks these two possible solutions. This explains the last condition of the theorem. □

**Proof of Theorem 3**.

As previously, we can assume $x \in (1/2, 1)$. The proof of Theorem 2 is immediately adapted if $x \geq y$, so that we focus on the case $x < y$. Using exactly the same computations as in the proof of Theorem 2, we can show that

$$|xz - \circ_\nu (xz_h + \circ_{\nu:+1}(xz_\ell))| < 2^{-2n-1} + \frac{1}{2}\mathrm{ulp}\left(xz_h + \circ_{\nu:+1}(xz_\ell)\right).$$

and Property 6 implies that we get a correctly rounded quotient.

# References


[1] C. Iordache and D. W. Matula. On infinitely precise rounding for division, square root, reciprocal and square root reciprocal. *Proc. 14th IEEE Symposium on Computer Arithmetic (Adelaide, Australia)*, pages 233–240, April 1999. IEEE Computer Society Press.

Table 1: Maximal errors (in ulps) of the naive solution for various values of $n$.

| $n$ | $x$ | $y$ | Error (in ulps) $>$ |
|---|---|---|---|
| 32 | $\frac{4294868995}{2147483648}$ | $\frac{65535}{32768}$ | $1.4999618524452582589456$ |
| 53 | $\frac{268435449}{134217728}$ | $\frac{9007199120523265}{4503599627370496}$ | $1.4999999739229677997443$ |
| 64 | $\frac{18446744066117050369}{9223372036854775808}$ | $\frac{18446744067635550617}{9223372036854775808}$ | $1.4999999994316597271551$ |
| 113 | $\frac{288230376151711737}{144115188075855872}$ | $\frac{10384593717069655112945804582584321}{5192296858534827628530496329220096}$ | $1.4999999999999999757138$ |

Table 2: The $n$-bit numbers $y$ between 1 and 2 for which, for any $n$-bit number $x$, $\circ_\nu(x \times \circ_\nu(1/y))$ equals $\circ_\nu(x/y)$.

| $n$ | | | | |
|---|---|---|---|---|
| 7 | 1 | $\frac{105}{64}$ | | |
| 8 | 1 | $\frac{151}{128}$ | $\frac{163}{128}$ | $\frac{183}{128}$ |
| 9 | 1 | $\frac{307}{256}$ | | |
| 10 | 1 | | | |
| 11 | 1 | $\frac{1705}{1024}$ | | |
| 12 | 1 | | | |
| 13 | 1 | $\frac{4411}{4096}$ | $\frac{4551}{4096}$ | $\frac{4915}{4096}$ |

Table 3: Number $\gamma(n)$ and percentage $100\gamma(n)/2^{n-1}$ of values of $y$ for which Algorithm 2 returns a correctly rounded quotient for all values of $x$. For $n \leq 7$, the algorithm always works.

| $n$ | $\gamma(n)$ | percentage |
|---|---|---|
| 7 | 64 | 100 |
| 8 | 127 | 99.218 |
| 9 | 254 | 99.218 |
| 10 | 510 | 99.609 |
| 11 | 1011 | 98.730 |
| 12 | 2022 | 98.730 |
| 13 | 4045 | 98.754 |
| 14 | 8097 | 98.840 |
| 15 | 16175 | 98.724 |
| 16 | 32360 | 98.754 |
| 17 | 64686 | 98.703 |
| 18 | 129419 | 98.738 |
| 19 | 258953 | 98.782 |
| 20 | 517591 | 98.722 |
| 21 | 1035255 | 98.729 |
| 22 | 2070463 | 98.727 |
| 23 | 4140543 | 98.718 |
| 24 | 8281846 | 98.727 |
| 25 | 16563692 | 98.727 |
| 26 | 33126395 | 98.724 |
| 27 | 66254485 | 98.726 |
| 28 | 132509483 | 98.727 |
| 29 | 265016794 | 98.726 |

[2] T. Lang and J.-M. Muller. Bound on run of zeros and ones for algebraic functions. *Proc. 15th IEEE Symposium on Computer Arithmetic (Vail, Colorado)*. IEEE Computer Society Press, 2001.

[3] M. A. Cornea-Hasegan, R. A. Golliver, and P. W. Markstein. Correctness proofs outline for Newton-Raphson based floating-point divide and square root algorithms. *Proc. 14th IEEE Symposium on Computer Arithmetic (Adelaide, Australia)*, pages 96–105, April 1999. IEEE Computer Society Press.

[4] P. W. Markstein. Computation of elementary functions on the IBM Risc System/6000 processor. *IBM Journal of Research and Development*, 34(1):111–119, Jan. 1990.

[5] C. Batut, K. Belabas, D. Bernardi, H. Cohen and M. Olivier, *User's Guide to PARI-GP*, available from `ftp://megrez.math.u-bordeaux.fr/pub/pari`.