Algorithms for triple-word arithmetic
Nicolas Fabiano, Jean-Michel Muller, Joris Picot

To cite this version:
Nicolas Fabiano, Jean-Michel Muller, Joris Picot. Algorithms for triple-word arithmetic. IEEE Transactions on Computers, Institute of Electrical and Electronics Engineers, In press, pp.1-11. 10.1109/TC.2019.2918451. hal-01869009v2

HAL Id: hal-01869009
https://hal.archives-ouvertes.fr/hal-01869009v2
Submitted on 20 May 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Algorithms for triple-word arithmetic
Nicolas Fabiano and Jean-Michel Muller and Joris Picot

1 INTRODUCTION AND NOTATION
Double-word and Triple-word arithmetics consist in representing a real number as the unevaluated sum of two and three floating-point numbers, respectively. They are frequently called “double double” and “triple double”, because in practice the underlying format being used is the binary64/double precision format of the IEEE 754 Standard for Floating-Point Arithmetic [7], [21].

A generalization of these arithmetics is the notion of floating-point expansion [26], [29], [10], where a high-precision number is represented as the unevaluated sum of \( n \) floating-point numbers.

Such arithmetics are useful. Numerical computations sometimes require a precision significantly higher than the one offered by the basic floating-point formats. Double-word, triple-word (or even quadruple-word) arithmetics have been used for implementing BLAS [18], [20], [30], for Semidefinite programming [22]. Bailey, Barrio and Borwein [1] give several timely examples in mathematical physics and dynamics where precision higher than double precision/binary64 are needed.

Another example is when evaluating transcendental functions with correct rounding: it is almost impossible to guarantee last-bit accuracy in the final result if all intermediate calculations are done in the target precision. For instance, the CRLibm library [3] of correctly rounded elementary functions uses double-word and triple-word [14], [15] operations in the last steps of the evaluation of approximating polynomials. The reason is simple: results on the table maker’s dilemma [16] show that returning a correctly rounded exponential or logarithm requires approximating the considered function with roughly twice the target precision, which in turn requires doing intermediate calculations with significantly more than twice the target precision. The Metalibm Lutetia library,\(^1\) also uses double-word and triple-word arithmetics.

Compared to double-word or triple-word arithmetic, arbitrary precision libraries such as GNU-MPFR [5] have the advantage of being versatile, but may involve a significant penalty in terms of speed and memory consumption if one only requires computations accurate within around 150 bits in a few critical parts of a numerical program.

Algorithms for double-word arithmetic have been presented in [17], [9]. The purpose of this paper is to introduce and analyze efficient algorithms for performing the arithmetic operations in triple-word arithmetic. Our goal is to obtain algorithms that are faster than the ones we could obtain simply by using floating-point expansion algorithms in the particular case \( n = 3 \), for a comparable accuracy.

In the following, we assume a radix-2, precision-\( p \) floating-point (FP) arithmetic system, with unlimited exponent range and correct rounding. As a consequence, our results will apply to “real-world” binary floating-point arithmetic, such as the one specified by the IEEE 754-2008 Standard, provided that underflow and overflow do not occur. We also assume the availability of an FMA (fused multiply-add) instruction. Such an instruction evaluates expressions of the form \( ab + c \) with one final rounding only.

The notation \( a | b \) means “\( a \) divides \( b \)”.

The notation RN(\( t \)) stands for \( t \) rounded to the nearest FP number, ties-to-even, and RU(\( t \)) (resp. RD(\( t \))) stands for \( t \) rounded towards \( +\infty \) (resp. \( -\infty \)). We will use three classical functions of the floating-point literature: ufp (unit in the last place [12], [21]), ufp (unit in the first place [28]) and uls (unit in the last significant place). They can be defined as follows. If \( x \neq 0 \) is a real number, then:

- \( \text{ufp}(x) = 2 \log_2 |x|; \)
- \( \text{ulp}(x) = \text{ufp}(x) \cdot 2^{-p+1}; \)
- \( \text{uls}(x) \) is the largest power of 2 that divides \( x \), i.e., the largest \( 2^k (k \in \mathbb{Z}) \) such that \( x/2^k \) is an integer.

When \( x \) is a FP number, \( \text{ufp}(x) \) is the weight of its most

\(^{1}\) See \url{http://www.metalibm.org/lutetia.html}
significant bit, $\text{ulp}(x)$ is the weight of its least significant bit, and $\text{uls}(x)$ is the weight of its rightmost nonzero bit. For instance, assuming binary64/double precision arithmetic ($p = 53$), if

$$x = -1.01101_2 \times 2^{364}$$

then $\text{ulp}(x) = 2^{364}$, $\text{ulp}(x) = 2^{312}$, and $\text{uls}(x) = 2^{359}$.

The number $u = 2^{-p} = 1/2\text{ulp}(1)$ denotes the roundoff error unit. When an arithmetic operation $a \lor b$ is performed, where $a$ and $b$ are FP numbers, what is effectively computed is $\text{RN}(a \lor b)$, and if $t$ is a real number and $T = \text{RN}(t)$, then

$$|t - T| \leq \frac{u}{1+u} |t| \leq u |t| \text{ and } |t - T| \leq u |T|,$$

and more precisely

$$|t - T| \leq \frac{1}{2} \text{ulp}(t) \leq \frac{1}{2} \text{ulp}(T).$$

The algorithms presented in this paper (as well as the usual algorithms that manipulate double-words or general expansions) use as basic blocks the classical algorithms Fast2Sum (Algorithm 1), 2Sum (Algorithm 2), and 2Prod (Algorithm 3) given below. Roughly speaking (more detail below), Algorithms 1 and 2 compute the error of a FP addition, and Algorithm 3 computes the error of a FP multiplication.

**Algorithm 1 – Fast2Sum**($a, b$). (3 operations) [4].

**Require:** $3$ integers $k_a \geq k_b, M_a, M_b$ (with $|M_a|, |M_b| \leq 2^p - 1$), s.t. $a = M_a \cdot 2^{k_a}$ and $b = M_b \cdot 2^{k_b}$.

**Ensure:** $s + e = a + b$

```latex
s \leftarrow \text{RN}(a + b)
\]
```

```
\]
```

Algorithm 2 requires twice as many operations as Algorithm 1, but its output variables always satisfy $s + e = a + b$: no knowledge of the respective magnitudes of $|a|$ and $|b|$ is needed.

**Algorithm 2 – 2Sum**($a, b$). (6 operations) [19], [13].

**Ensure:** $s + e = a + b$

```latex
s \leftarrow \text{RN}(a + b)
\]
```

The values $\pi$ and $e$ computed by Algorithm 3 satisfy $\pi + e = ab$. Algorithm 3 uses an FMA instruction (for computing $\text{RN}(a \cdot b - \pi)$).

Sometimes, we know in advance the value of $s$ or $\pi$. In that case, $e$ can be computed saving the first operation, with algorithms denoted in the following by for instance $2\text{Sum}_2(s, y)$.

When one defines a number as the unevaluated sum of two, three or more FP numbers, one has to explain to which extent they can “overlap”: after all the sum of the three double-precision/binary64 numbers 1, 2, and 4 is just a three-bit number, expressing it as the sum of three FP numbers does not make it more accurate. Several definitions appear in the literature. The first needed in this paper is Priest’s definition:

**Definition 1.** The sequence $(x_i)$ is P-nonoverlapping (with Priest’s definition [27]) when $\forall i, |x_{i+1}| < \text{ulp}(x_i)$.

We also introduce the following definition, more restrictive than Shechcuk’s definition [29]:

**Definition 2.** The sequence $(x_i)$ is F-nonoverlapping (with Fabiano’s definition) when $\forall i, |x_{i+1}| \leq \frac{1}{2} \text{uls}(x_i)$.

**Definition 3.** For any definition of nonoverlapping, a sequence $(x_i)$ is said nonoverlapping if (with possible interleaving zeros) when we have a set $I_0$ such that $\forall i \in I_0, x_i = 0$, and $(x_i)_{i \notin I_0}$ is nonoverlapping.

We can now formally define the double-word and triple-word numbers:

**Definition 4.** We call double-word number (DW) a pair $(x_0, x_1)$ of FP numbers such that $x_0 = \text{RN}(x_0 + x_1)$.

**Definition 5.** We call triple-word number (TW) a triplet $(x_0, x_1, x_2)$ of FP numbers that is P-nonoverlapping.

Note the deliberate difference between Definitions 4 and 5. To obtain a definition similar to Definition 4, one could define a triple-word number as a 3-uple $(x_0, x_1, x_2)$ of floating-point numbers such that $x_0 = \text{RN}(x_0 + x_1 + x_2)$ and $x_1 = \text{RN}(x_1 + x_2)$. Such a requirement would be much stronger than Definition 5, resulting in practice in more complex and less efficient algorithms.

The definition of general expansions in [25] is based on ulp-nonoverlapping ($\forall i, |x_{i+1}| \leq \text{ulp}(x_i)$), which is slightly less restrictive than the one we chose for TW. Therefore algorithms proven for general expansions may not be correct for TW.

The paper is organized as follows. Section 2 presents some other basic blocks that will be used in the rest of the paper, and some original results related to them. Section 3 proves that the classical algorithm to turn
an arbitrary sequence into an expansion works for TW. Section 4 presents an algorithm to correctly round a TW to a FP number, and proves its correctness. Section 5 does the same as section 3 for the sum of two TW. Section 6 presents two versions of an original algorithm for computing the product of two TW, proves their correctness and gives tight error bounds. Sections 7 to 10 provide a similar analysis, in the case of the product of a DW by a TW, the reciprocal of a TW, the quotient of two TW, and the square root of a TW, respectively. Section 11 compares our results to the ones known for general n-word expansions, with n = 3.

2 OTHER BASIC BLOCKS

The Algorithms on TW presented in this paper use as basic blocks the 2Sum, Fast2Sum and 2Prod algorithms presented in the previous section, as well as the following, less classical, VecSum and VecSumErrBranch algorithms. Many properties of these algorithms have been proven elsewhere [26], [24], [2], but in this paper, we will need specific properties, presented below.

2.1 VecSum

The VecSum algorithm (Algorithm 4) first appears as a part of Priest’s normalization algorithm [26]. The name “VecSum” was coined by Ogita et al [24]. The aim of this algorithm is to turn a sequence that is “slightly” nonoverlapping into one that is “more” nonoverlapping, with no error. It is illustrated Fig. 1.

Algorithm 4 – VecSum(x₀, . . . , xₙ₋₁). (6n – 6 operations)

Ensure: e₀ + · · · + eₙ₋₁ = x₀ + · · · + xₙ₋₁

s₀ ← x₀
for i = n – 2 to 0 do
sᵢ, eᵢ₊₁ ← 2Sum(xᵢ, sᵢ)
end for

return (e₀, e₁, . . . , eₙ₋₁)

Fig. 1. The VecSum Algorithm [26], [24].

There are several theorems related to this algorithm, that use different definitions of nonoverlapping. In what follows, we will use the following original result:

Theorem 1. Assume, after removing possible interleaving zeros, that we can write in a non-necessarily canonical way (i.e., we do not necessarily have |Mᵢ| ≥ 2⁻ⁿ⁻¹) xᵢ = Mᵢ,2ᵏᵢ−pᵢ + 1, |Mᵢ| < 2ᵖᵢ, such that ∀i ≤ n – 2, kᵢ₋₁ ≥ kᵢ + 1, and kᵢ₋₂ ≥ kᵢ₋₁. Then VecSum(x₀, . . . , xₙ₋₁) is F-nonoverlapping with the same sum.

In this case, Fast2Sum can be used instead of 2Sum, so that VecSum only costs 3n – 3 operations.

Proof: Interleaving zeros in the input simply give some interleaving zeros in the output without changing the non-zero terms, so that we can suppose that we have removed them.

Firstly, ∀i, |sᵢ| ≤ (2 – 2u)2ᵏᵢ−₁. Indeed, if by induction |sᵢ₊₁| ≤ (2 – 2u)2ᵏᵢ, given |xᵢ| ≤ (2 – 2u)2ᵏᵢ, we get |sᵢ₊₁| + |xᵢ| ≤ (4 – 4u)2ᵏᵢ ≤ (2 – 2u)2ᵏᵢ−₁ so |sᵢ| ≤ (2 – 2u)2ᵏᵢ−₁.

This gives |eᵢ| ≤ 2u2ᵏᵢ−₁, and justifies Fast2Sum being used.

We suppose that |eᵢ| > 1uls(eᵢ′) with i′ < i. We also suppose without loss of generality that uls(eᵢ′) = u. We easily get by induction that for all i, if 2ᵏᵢ |sᵢ, xᵢ₋₁, . . . , x₀, then 2ᵏᵢ′ |eᵢ, . . . , e₀. Yet |eᵢ| ≤ 1₂ulp(sᵢ) so |sᵢ| ≥ 1 so sᵢ is a multiple of 2u. Given we want a eᵢ′ non-multiple of 2u, that must be the case for one of the xᵢ, j ≤ i – 2. In particular, we have 2ᵏᵢ ≤ 1₂, hence 2ᵏᵢ−₂ ≤ 1₂, so 2ᵏᵢ−₁ ≤ 1₂, which contradicts |eᵢ| ≤ 2u2ᵏᵢ−₁.

The conditions on the input of Theorem 1 are complex, so we will use the following corollary:

Corollary 1. Assume that we have I ⊆ [1, n − 2] with no 2 consecutive indices such that ∀i ∈ [0, n − 2], i ∉ I, ufp(xᵢ₊₁) ≤ 1₂ ufp(xᵢ), and ∀i ∈ I, ufp(xᵢ₊₁) ≤ 2⁻ⁿ⁻₂ulp(xᵢ) and ufp(xᵢ₊₁) ≤ 1₄ ufp(xᵢ₋₁).

Then VecSum(x₀, . . . , xₙ₋₁) is F-nonoverlapping with the same sum. In this case, Fast2Sum can be used instead of 2Sum, so that Algorithm 4 only costs 3n – 3 operations.

Proof: For i ∉ I, we take kᵢ = eᵢ the canonical exponent, and for i ∈ I, we take kᵢ = max(kᵢ₊₁ + 1, eᵢ₊₁). This is possible because: 2ᵏᵢ₊₁−pᵢ₊₂|xᵢ| and 2ₑᵢ₊₁−pᵢ₊₁|xᵢ|, which imply 2ᵏᵢ−pᵢ₊₁|xᵢ| and |xᵢ| ≤ 2·2ᵉᵢ, which imply |xᵢ| ≤ 2·2ᵏᵢ.

For i, i + 1 ∉ I, we have kᵢ₊₁ ≤ kᵢ − 1. For i ∈ I, we have on one hand kᵢ₊₁ ≤ kᵢ − 1, and on the other hand eᵢ₊₁ ≤ kᵢ−₁ − 1 and kᵢ₊₁ ≤ kᵢ−₁ − 2 so kᵢ ≤ kᵢ−₁ − 1.

2.2 VecSumErrBranch

VecSumErrBranch (Algorithm 5 below) has similarities with Algorithm 4, but sums are computed starting from the larger terms, and some tests help avoiding to return too many zero terms. It is illustrated Fig. 2. It can be traced back to a part of Hida, Li and Bailey’s normalization algorithm for “quad-double” numbers [6], which itself is a variant of Priest’s normalization algorithm [26].
Algorithm 5 – VSEB($e_0, \ldots, e_{n-1}$). ($6n - 6$ operations & $n - 2$ tests)

Ensure: $y_0 + \cdots + y_{n-1} = e_0 + \cdots + e_{n-1}$

\begin{algorithm}
\begin{algorithmic}
\State $j \leftarrow 0$
\State $\epsilon_0 \leftarrow e_0$
\For {$i = 0$ to $n - 3$}
\State $(r_i, \epsilon_{i+1}) \leftarrow 2\text{Sum}(e_i, \epsilon_{i+1})$
\If {$\epsilon_{i+1} \neq 0$}
\State $y_j \leftarrow r_i$
\State $\epsilon_{i+1} \leftarrow \epsilon_{i+1}$
\State incr $j$
\EndIf
\EndFor
\State $y_j, y_{j+1} \leftarrow 2\text{Sum}(\epsilon_{n-2}, \epsilon_{n-1})$
\State $y_{j+2}, \ldots, y_{n-1} \leftarrow 0$
\State return $(y_0, y_1, \ldots, y_{n-1})$
\end{algorithmic}
\end{algorithm}

Fig. 2. The VecSumErrBranch (VSEB) Algorithm.

We prove the following property, of Algorithm 5, that will be useful later on.

**Theorem 2.** If ($e_i$) is F-nonoverlapping wIZ, then VSEB($x_0, \ldots, x_{n-1}$) is P-nonoverlapping with the same sum, provided that $p \geq n - 1$. In this case, Fast2Sum can be used instead of 2Sum, so that Algorithm 5 only costs $3n - 3$ operations.

**Proof:** Again, interleaving zeros in the input can be ignored without changing anything, so we suppose that we have removed them. We write $e_i = M_i \cdot 2^{k_i}$ with $|M_i| < 2^p$ odd so that $|\epsilon_{i+1}| \leq \frac{1}{2}2^{k_i}$. By an easy induction, for all $i$, $r_{i-1}$ and $\epsilon_i$ are multiples of $2^{k_i}$.

- Let $i_0$ be such that $|\epsilon_{i_0}| = 2^{k_{i_0}}$. Let us show by induction that for all $i_0 \leq i \leq n - 2$, we have $|r_i| \leq 2^{k_{i_0}}(2 - 2^{i_0 - 1})$.
  - We can initialize with $i = i_0 - 1$, because what is transmitted to next step (playing the role of $r_{i_0}$) is $\epsilon_{i_0}$, which exactly satisfies the condition.
  - We suppose that $|r_{i-1}| \leq 2^{k_{i_0}}(2 - 2^{i_0 - 1})$.
    We have $|\epsilon_{i+1}| \leq \frac{1}{2}2^{k_i} = \frac{1}{2}2^{k_{i-1}} \leq \cdots \leq 2^{i_0 - 2}2^{k_{i_0}}$.
    Thus $|r_{i-1}| + |\epsilon_{i+1}| \leq 2^{k_{i_0}}(2 - 2^{i_0 - 1})$, and this is a FP number because $i_0 \geq 1$ and $i \leq n - 2$ give $i_0 - i - 1 \geq -n + 2 \geq -p + 1$, so that $|r_i| \leq 2^{k_{i_0}}(2 - 2^{i_0 - 1})$. This gives the result by induction.

In particular, with $j$ such that $y_j = r_{i_0 - 1}$, we have $\text{ulp}(y_j) \geq 2|\epsilon_{i_0}| = 2 \cdot 2^{k_{i_0}}$ so $|y_{j+1}| < \text{ulp}(y_j)$.

- For $i_0$ such that $|\epsilon_{i_0}| > 2^{k_{i_0}},$ we similarly get that for all $i \geq i_0$ such that $\epsilon_{i_0+1}, \ldots, \epsilon_i = 0,$ we have $|r_i| \leq |\epsilon_{i_0}| + 2^{k_{i_0}}$.

Indeed, the same proof replacing $2^{k_{i_0}}(2 - \cdots)$ by $|\epsilon_{i_0}| + 2^{k_{i_0}}(1 - \cdots)$ works, except that the equality case can be reached in case of errors. In this case, this is sufficient given $\text{ulp}(y_j) \geq 2|\epsilon_{i_0}| > |\epsilon_{i_0}| + 2^{k_{i_0}} \geq y_{j+1}$.

Finally, we can use Fast2Sum. Indeed, we have $2^{k_i} e_0, \ldots, e_i$ so $2^{k_i} |\epsilon_i|$, and $|\epsilon_{i+1}| \leq 2^{k_i}$.

\[\square\]

Usually, we do not want to keep the complete output, but only a fixed number of terms. The resulting algorithm is denoted by VSEB($k$)($e_0, \ldots, e_{n-1}$). We have

**Theorem 3.** The relative error caused by keeping only the first $k$ terms of a P-nonoverlapping sequence is bounded by $2u^k + 4.2 \cdot u^{k+1}$, provided that $p \geq 6$.

**Proof:** We have by P-nonoverlapping:

$$\text{ulp}(y_k) \leq u \cdot \text{ulp}(y_{k-1}) \leq \cdots \leq u^k \cdot \text{ulp}(y_0),$$

due to,

$$|y_k| + \cdots + |y_{n-1}| \leq (2 - 2u)(u^k + u^{k+1} + \cdots + u^p)\text{ulp}(y_0),$$

due to $|y_k + \cdots + y_{n-1}| \leq 2u^k \text{ulp}(y_0)$.

We also have,

$$|y_0 + \cdots + y_{n-1}| \geq |y_0| - |y_1 + \cdots + y_{n-1}| \geq (1 - 2u)\text{ulp}(y_0).$$

Therefore,

$$|y_k + \cdots + y_{n-1}| \leq \frac{2u^k}{1 - 2u}|y_0 + \cdots + y_{n-1}|,$$

which implies the theorem. \[\square\]

### 2.3 Composed algorithm

Algorithms VecSum and VSEB were designed to be composed, in order to obtain a “normalization” algorithm such as Algorithm 6 below. In that case, we note that the first 2Sum in VSEB can be skipped because ($e_0, e_1$) is already a DW. When Fast2Sum can be used everywhere, we get a total of $6n - 9$ operations and $n - 2$ tests.

### 3 Arbitrary three FP numbers to TW

Before manipulating Triple-Word numbers, we want to be able to turn any unevaluated sum of three FP numbers into a TW, with no error (this is the equivalent of the 2Sum algorithm for 3 FP numbers). We can use Algorithm 6, which can be found in [6] for quad-word numbers, and in [25, page 87] for general expansions:
We denote $s$

Algorithm 6 – ToTW($a$, $b$, $c$). (21 operations & 1 test)

Ensure: $\bar{r}$ TW and $\bar{r} = a + b + c$

d₀, d₁ ← 2Sum($a$, $b$)
e₀, e₁, e₂ ← VecSum(d₀, d₁, $c$

Algorithm 7 – RoundTW($x₀$, $x₁$, $x₂$). (3 operations & 4 tests)

Require: $\bar{x}$ TW

Ensure: $y \in \text{RN}(\bar{x})$

if $\text{RN}(x₀ + 2x₁)$ inexact operation or $\star$ $\text{RN}(-\frac{3}{2}u - 2u^2) \cdot x₀ \neq x₁$ then

$y \leftarrow \text{RN}(x₀ + x₁)$

else if $x₂ > 0$ then

$y \leftarrow \text{RU}(x₀ + x₁)$

else if $x₂ < 0$ then

$y \leftarrow \text{RD}(x₀ + x₁)$

else

$y \leftarrow \text{RN}(x₀ + x₁)$

end if

return $y$

We have,

Theorem 4. If $a$, $b$, $c$ are FP numbers, then $\text{ToTW}(a, b, c)$ is a TW, provided that $p \geq 4$.

Proof: If $(e₀, e₁, e₂)$ is F-nonoverlapping, then Theorem 2 concludes.

First, $|e₁| \leq \frac{1}{2}\text{ulp}(e₀)$ gives $(e₀, e₁)$ F-nonoverlapping.

We denote $s := \text{RN}(d₁ + z)$ the intermediate value in VecSum($d₀, d₁, c$).

- If $e₁ \neq 0$, we suppose without loss of generality that $\text{ulp}(e₁) = u$, and $e₂ \geq \frac{1}{2}u$. Then $|e₂| \leq \frac{1}{2}\text{ulp}(s)$ gives $s \geq 1$ so $2u$ is but $e₁$ is not divisible by $2u$ so $d₀$ neither, hence $d₀ < 1$, so $|d₁| \leq \frac{1}{2}\text{ulp}(d₀) \leq \frac{1}{2}u$.

Furthermore, $|c + d₁| \geq 1 + \frac{1}{2}u$. Thus $|c| \geq (1 + \frac{1}{2}u) - \frac{1}{2}u = 1 \text{ so } \text{ulp}(c) \geq 2u \geq 2|d₁|$. So $|e₁| = \frac{1}{2}u \geq |d₁|$, which is impossible since $|e₂| \geq \frac{1}{2}u \geq |d₁|$.

So $(e₀, e₁)$ is F-nonoverlapping too.

- If $e₀ = 0$, then the same reasoning works with $e₀$ instead of $e₁$.

Another typical way of forming a TW consists in using any of the following algorithms, but with inputs that are DW instead of TW (with an implicit third term equal to zero, and simplifications obtained by removing useless operations).

4 Rounding a TW to a FP Number

Frequently, TW arithmetic is used in intermediate calculations, but the final result must be a floating-point number. Hence, we need to be able to return the FP number closest to a TW number. This is for typically the case when we use TW numbers in intermediate calculations for implementing correctly rounded elementary functions. This can be done using Algorithm 7.

At line 1 of Algorithm 7, although the fact that the operation $x₀ + 2x₁$ is exact can in theory be detected by checking a flag, it will be more efficient in practice to compute $(s, e) = \text{Fast2Sum}(x₀, 2x₁)$ and test whether $e$ is zero.

We have,

Theorem 5. If $\bar{x}$ is a TW, then $\text{RoundTW}(\bar{x}) = \text{RN}(\bar{x})$, provided that $p \geq 4$.

Proof: First, if $x₀ + x₁$ is a FP number, then $y = x₀ + x₁$ anyway, and it is easy to check that $x₀ + x₁ = \text{RN}(\bar{x})$. We suppose for the sequel of the proof that this is not the case.

Given $|x₁| < \text{ulp}(x₀)$, the first condition is false iff $x₀ + x₁$ is halfway between two consecutive FP numbers, or in a special case that can without loss of generality be reduced to $x₀ = 1 + 2u$ and $x₁ = -\frac{3}{2}u$. When that first condition is false, Condition $(\star)$ is designed to be true in the special case, but false elsewhere (because of the magnitude of $|x₁|$).

- If $x₀ + x₁$ is halfway between two adjacent FP numbers, then the rounding is decided by the sign of $x₂$.

- Otherwise, one easily checks that $\text{RN}(x₀ + x₁ + x₂) = \text{RN}(x₀ + x₁)$, given $|x₂| < \text{ulp}(x₁)$.

5 Sum of two TW Numbers

To compute the sum of two TW numbers, we simply use the composition of VecSum and VSEB after a preliminary sorting of the input. This gives Algorithm 8 below. That algorithm would of course also work to compute the sum of a DW and a TW with the same error bound, but with less operations and tests.

Algorithm 8 – TWSum($x₀, x₁, x₂, y₀, y₁, y₂$). (42 operations & 8 tests)

Require: $\bar{x}$ and $\bar{y}$ TW

Ensure: $\bar{r}$ TW and $\left| \frac{f(\bar{x} + \bar{y})}{\bar{x} + \bar{y}} \right| \leq 2u^3 + 4.2u^4$

$z_0, \ldots, z_5 \leftarrow \text{Merge}\left(\begin{array}{l} (x₀, x₁, x₂), (y₀, y₁, y₂) \\ (e₀, \ldots, e₅) \end{array}\right)$

e₀, \ldots, e₅ ← VecSum($z₀, \ldots, z₅$

$r₀, r₁, r₂ ← \text{VSEB}(3)(e₀, \ldots, e₅)$

return $(r₀, r₁, r₂)$

5.1 Correctness of Algorithm 8

We have,

Theorem 6. Let $x₀, \ldots, x₅$ be FP numbers such that

$∀ i, |x_{i+1}| \leq |x_i| \text{ and } ∀ i, |x_{i+2}| < \text{ulp}(x_i)$

Then $\text{VSEB}(\text{VecSum}(x₀, \ldots, x₅))$ is P-nonoverlapping, provided that $p \geq 4$.

Interestingly enough, we can notice that Theorem 6 may not hold for more than 6 floating-point inputs. Indeed, for 7 inputs, we can consider

$(x_i) = 1-u, -1+2u, -u+u^2, u-u^2, u^2-u^3, u^2-u^3, u^3-u^4$
which gives \((e_i) = u, u^2, u^3, -u^4, \) and finally
\[(y_i) = u, 2u^2, -u^3, -u^4\]
with \(2u^2 = \text{ulp}(u)\).

This is why it is reasonable to use the notion of P-nonoverlapping for TW numbers only, but not for
general expansions, for which Algorithm 8 preserves ulp-nonoverlapping only [25, page 90].

Sketch of the proof: For space constraints, the proof
of Theorem 6 is not detailed. The main steps are:
• prove by induction that \(|s_i| \leq 2\text{ulp}(x_{i-1})\) and \(|s_i| \leq 4\text{ulp}(x_i)\);
• if \(e_i > \frac{1}{2}\text{uls}(e_j)\) for some \(j < i\), deduce some
  conditions on \(i\) and the nearby terms in various
  cases;
• conclude with a case study: \(i \leq 3\) and \(e_i > \frac{1}{2}\text{uls}(e_j)\); \(i \geq 4\) and \(e_i > \frac{1}{2}\text{uls}(e_j)\); or \(0 < e_i \leq \frac{1}{2}\text{uls}(e_j)\).

5.2 Number of operations in Algorithm 8

In the Merge instruction of Algorithm 8, if the last two
numbers to sort are \(x_2\) and \(y_2\), there is no need to do
it because they play symmetrical roles in 2Sum. Thus
this part costs only 4 tests. In VecSum, there are for each
block examples where Fast2Sum cannot be used, so it
costs 30 operations. One easily checks that Fast2Sum
can be used in VSEB, so it costs 12 operations and 4 tests.

6 PRODUCT OF TWO TW NUMBERS

To compute the product of two TW numbers, we simply
distribution the sub-products and aggregate the terms
ensuring P-nonoverlapping, with an error as small as
possible. The algorithms presented below guarantee commutativity, even if it is rarely useful in practice.

Algorithm 9 – 3Prod\(t\)(\(x_0, x_1, x_2, y_0, y_1, y_2\)). (46 operations & 2 tests)

Require: \(\bar{x}\) and \(\bar{y}\) TW ; \(p \geq 6\)
Ensure: \(\bar{z}\) TW and \(\frac{\bar{x} \cdot \bar{y} - \bar{z}}{\text{ulp}(x)} \leq 28u^3 + 107u^4\)
\begin{align*}
  & z^0_{00}, z_{00}^0 \leftarrow 2\text{Prod}(x_0, y_0), \quad z_{00}^+ = 2\text{Prod}(x_0, y_0) \\
  & z^0_{01}, z_{01}^0 \leftarrow 2\text{Prod}(x_0, y_1), \quad z_{01}^+ = 2\text{Prod}(x_1, y_0) \\
  & z^0_{10}, z_{10}^+ \leftarrow 2\text{Prod}(x_1, y_0) \\
  & b_0, b_1, b_2 \leftarrow \text{VecSum}(z^0_{00}, z_{01}^0, z_{10}^+). \\
  & c \leftarrow \text{RN}(b_2 + x_1y_1) \quad \text{(FMA)} \\
  & z_{31} \leftarrow \text{RN}(z^+_{10} + x_0y_2) \quad \text{(FMA)} \\
  & z_{32} \leftarrow \text{RN}(z_{01}^+ + x_2y_0) \quad \text{(FMA)} \\
  & z_3 \leftarrow \text{RN}(z_{31} + z_{32}) \quad \text{(FMA)} \\
  & e_0, e_1, e_2, c, e_4 \leftarrow \text{VecSum}(z^0_{00}, b_0, c, e_3) \\
  & r_0 \leftarrow e_0 \\
  & r_1, r_2 \leftarrow \text{VSEB}(2)(e_1, e_2, c, e_4) \\
\end{align*}
return \((r_0, r_1, r_2)\)

6.1 Bounds on the different terms

We suppose without loss of generality that \(1 \leq x_0, y_0 < 2\),
so that \(|x_1|, |y_1| < 2u\) and \(|x_2|, |y_2| < 2u^2\). Then, we have:
\[
\begin{align*}
  1 & \leq |z^+_{00}| < 4 \\
  |z^0_{00}| & \leq 2u : \text{uls}(z^0_{00}) \geq 4u^2 \\
  |z^0_{10}|, |z^0_{10}| & < 4u : |b_2| \leq 4u^2 \\
  |x_1y_1| & < 4u^2 - 4u^3 : |c| < 8u^2 \\
  |z^0_{11}|, |z^0_{10}| & \leq 2u^2 : |x_0y_0|, |x_2y_0| \leq 4u^2 \\
  |z_{31}, z_{32}| & \leq 6u^2 : |z_3| \leq 12u^2 \\
  |s_3| & \leq 20u^2 \\
\end{align*}
\]

6.2 Correctness and error bound of Algorithm 9

We have,

Theorem 7. If \(\bar{x}, \bar{y}\) are TW numbers and \(p \geq 6\), then
3Prod\(t\)(\(x, y\)) is a TW number, and the relative error
committed by 3Prod\(t\)(\(x, y\)) is bounded by \(28u^3 + 107u^4\).

Theorem 7 uses the following, straightforward, lemma.

Lemma 1. For all FP numbers \(x, y, \frac{1}{2}\text{ulp}(x)\text{RN}(x + y)\).

Proof of the theorem:
1. \((r_0, r_1, r_2)\) is a TW number.
2. First, let us prove that the last 2 lines are equivalent
to computing \(r_0, r_1, r_2 = \text{VSEB}(3)(e_0, e_1, e_2, e_3, e_4)\)
• If \(e_1 \neq 0\), then the fact that \(e_0 = \text{RN}(e_0 + e_1)\)
  concludes immediately.
• If \(e_1 = 0\), one easily checks that \(|s_1|, |s_2|, |s_3| < 16u \leq \frac{1}{2}\text{ulp}(z^+_{00})\) so that the next nonzero \(|e_i|\) is strictly less
  than \(\frac{1}{2}\text{ulp}(e_0)\), which concludes.

3. Then, let us prove that, with this equivalent version,
   \((r_0, r_1, r_2)\) is P-nonoverlapping.

We denote \(a := \text{RN}(z^+_{00} + z^+_{10})\) and \(s_3 := \text{RN}(c + z_3)\)
intermediate sums in VecSum.

We want to show that VecSum\(z^+_{00}, b_0, b_1, s_3\) is F-
nonoverlapping, with \(e_4\) F-nonoverlapping them too.

Thanks to Theorem 2, this would imply that \((r_0, r_1, r_2)\)
is P-nonoverlapping.

- First, let us show that \((z^+_{00}, b_0, b_1, s_3)\) satisfies the
  conditions of Theorem 1.

  First, we always have \(\text{ulp}(z^{00}_0) \geq 1\) much larger than
  four times any other number computed ; and in case on
  they are non-zero \(\text{ulp}(b_1) \leq \frac{1}{2}\text{ulp}(b_0) < \frac{1}{2}\text{ulp}(b_0)\).

  For the rest of the proof, we suppose without loss of
generality that \(|x_1| \geq |y_1|\).

On one hand, we easily obtain \(|s_3| \leq 10\text{ulp}(x_1)\).

On the other hand lemma 1 gives \(\frac{1}{2}\text{ulp}(x_1)\|a\text{ with}
\frac{1}{2}\text{ulp}(x_1) < u^2 < \text{uls}(z^{00}_0)\text{ so }\frac{1}{2}\text{ulp}(x_1)|b_0, b_1|.

• Case \(s_3 = 0\): we use \(I = 0\). (nothing more to show)
• Case \(s_3 \neq 0\) and \(b_0 = 0\): we use \(I = 0\) (idem)
• Case \(s_3 \neq 0, b_0 \neq 0\) and \(b_1 = 0\): we use \(I = \{1\}.\)
Indeed, we have $\text{ufp}(s_3) \leq \frac{1}{2} \text{ufp}(z_{00})$, and $\text{ufp}(s_3) \leq \frac{1}{2} 2^{p-2}\text{uls}(b_0)$, (using $p \geq 6$).

- Case $s_3 \neq 0$, $b_0 \neq 0$ and $b_1 \neq 0$: we use $I = \{2\}$. Indeed, we have $\text{ufp}(s_3) \leq 16\text{ufp}(b_1) \leq \frac{1}{2} \text{ufp}(b_0)$ and $\text{ufp}(s_3) \leq \frac{1}{2} 2^{p-2}\text{uls}(b_1)$, (using $p \geq 6$).

It also allows us to call Fast2Sum instead of the three 2Sum in this part of the algorithm, saving FP operations.

- Then, let us show that $e_4$ is F-nonoverlapping with the other $e_i$.

Firstly, $\text{ulp}(s_3) \geq 2|e_4|$. Secondly, we have seen that $\text{uls}(b_0), \text{uls}(b_1) \geq \frac{1}{2} \text{ulp}(x_i) \geq \frac{1}{20}|s_3| \geq 2\text{ulp}(s_3)$. (using $p \geq 6$). Thirdly, $\text{ulp}(z_{00}^+ \geq 2u > \text{ulp}(s_3)$.

Thus $e_0, e_1$ and $e_2$ are divisible by $\text{ulp}(s_3) \geq 2|e_4|$.

### 2. The relative error is bounded by $28u^3 + 107u^4$.

There are three sources of error: the terms that are ignored, the roundings in the computation of $z_3$ and $c$ and the terms not kept in VSEB. A naive analysis gives:

$$|e_0| = |x_1y_2 + x_2y_1 + x_2y_2| 
\leq (2(2u - 2u^2)(2u^2 - 2u^3) + (2u^2 - 2u^3)^2 
\leq 8u^3 - 11.9u^4$$

$$|e_1| = |(z_{10}^+ + x_0y_2) - z_{3,1}| 
\leq u \cdot \text{ulp}(z_{10}^+ + x_0y_2) 
\leq u \cdot \text{ulp}(2u^2 + 4u^2) 
\leq 4u^3$$

$$|e_2| = |(z_{01}^+ + x_2y_0) - z_{3,2}| \leq 4u^3$$

$$|e_3| = |(z_{3,1} + z_{3,2}) - z_3| \leq 8u^3$$

$$|e_4| = |(b_2 + x_1y_1) - c| \leq 4u^3$$

$$|e_5| = |(z_{00}^+ + b_0 + b_1 + c + z_3) - (r_0 + r_1 + r_2)| 
\leq (2u^3 + 4.2u^4)|z_{00}^+ + b_0 + b_1 + c + z_3|$$

Yet, $x, y \geq 1 - (2u - 2u^2)^2 - (2u^2 - 2u^3) \geq 1 - 2u$ so $x_0y \geq 1 - 2u$. We eventually obtain that the error cannot be too large if $e_0 \neq 0$ (not detailed here), and finally:

$$\frac{|\bar{r} - \bar{x}\bar{y}|}{\bar{x}\bar{y}} \leq \frac{28u^3 - 11.9u^4}{1 - 2u} \leq 28u^3 + 107u^4.$$  \hfill (1)

The bound (1) is very tight. Indeed, for instance in binary64 arithmetic, if we take:

$$(x_0, x_1, x_2) = (1 + 13 \cdot 2^{26} + 28)u, 2u - 2^{27}u^2, 2u^2 - 4u^3)$$

$$(y_0, y_1, y_2) = (1 + 7 \cdot 2^{27}u, 2u - (2^{28} - 8)u^2, 2u^2 - 4u^3),$$

we obtain a relative error around $(28 - 10^{-5})u^3$.

### 6.4 Faster version

In this version, $e_4$ is not computed.

**Algorithm 10 – 3Prod$^{\text{fast}}_{3,3} (x_0, x_1, x_2, y_0, y_1, y_2)$. (38 operations & 1 test)**

**Require:** $\bar{x}$ and $\bar{y}$ TW ; $p \geq 6$

**Ensure:** $\bar{r}$ TW and $\left| \frac{\bar{r}}{\bar{x}\bar{y}} \right| \leq 44u^3 + 176u^4$

[same 5 first lines as Algorithm 9]

$z_{3,1} \leftarrow \text{RN}(z_{10}^+ + x_0y_2)$ (FMA)

$z_{3,2} \leftarrow \text{RN}(z_{01}^+ + x_2y_0)$ (FMA)

$z_3 \leftarrow \text{RN}(z_{3,1} + z_{3,2})$

$s_3 \leftarrow \text{RN}(c + z_3)$

$e_0, e_1, e_2, e_3 \leftarrow \text{VecSum}(z_{00}^+, b_0, b_1, s_3)$

$r_0 \leftarrow e_0$

$r_1, r_2 \leftarrow \text{VSEB}(2)(e_1, e_2, e_3)$

return $(r_0, r_1, r_2)$

This saves 8 operations and 1 test. The proof of correctness and the computation of the error bound are similar to those for Algorithm 9. The obtained error bound is $44u^3 + 176u^4$, which is very tight, because with the input values (2) in binary64 arithmetic, we obtain a relative error around $(44 - 10^{-5})u^3$.

### 7 Multiplying a DW by a TW

To multiply a DW $(x_0, x_1)$ by a TW $(y_0, y_1, y_2)$, we use the same algorithms as previously, slightly simplified to take into account that $x_2 = 0$. Additionally to being interesting in itself, this operation will be used later on to compute reciprocals and quotients, which justifies a separate analysis.

**Algorithm 11 – 3Prod$^{\text{fast}}_{3,3} (x_0, x_1, y_0, y_1, y_2)$. (45 operations & 2 tests)**

**Require:** $\bar{x}$ DW and $\bar{y}$ TW ; $p \geq 6$

**Ensure:** $\bar{r}$ TW and $\left| \frac{\bar{r}}{\bar{x}\bar{y}} \right| \leq 10.5u^3 + 39u^4$

[same 5 first lines as Algorithm 9]

$z_{3,1} \leftarrow \text{RN}(z_{10}^+ + x_0y_2)$ (FMA)

$z_3 \leftarrow \text{RN}(z_{3,1} + z_0^+)$

$e_0, e_1, e_2, e_3, e_4 \leftarrow \text{VecSum}(z_{00}^+, b_0, b_1, c, z_3)$

$r_0 \leftarrow e_0$

$r_1, r_2 \leftarrow \text{VSEB}(2)(e_1, e_2, e_3, e_4)$

return $(r_0, r_1, r_2)$

### 7.1 Correctness, number of operations and error bound

Since Algorithm 11 is a particular case of Algorithm 9, correctness is directly ensured, provided that $p \geq 6$. Compared to Algorithm 9, we saved 1 operation, so that the total number of operations is 41. The error bound of Theorem 7 holds. However, one can redo the analysis taking into account that $x_2 = 0$ and obtain a better bound, namely,
Theorem 8. If \( \bar{x} \) is a DW number and \( \bar{y} \) is a TW number, then the relative error committed by \( 3\text{Prod}_{2,3}^{\text{ac}}(\bar{x}, \bar{y}) \) is bounded by \( 10.5u^3 + 39u^4 \), provided that \( p \geq 6 \).

The proof is omitted. It is unclear whether the bound of Theorem 8 is very tight, but it is not too bad either: in binary64 arithmetic, the choice

\[
\begin{align*}
(x_0, x_1) &= (1 + 3 \cdot 2^{-27} u, u - 2^{-27} u^2) \\
(y_0, y_1, y_2) &= (1 + (3 \cdot 2^{26} + 6)u, 2u - 5 \cdot 2^{27} u^2, 2u^2 - 26u^3)
\end{align*}
\]

(3)

gives a relative error around \( (10 - 2 \cdot 10^{-6})u^3 \).

7.2 Faster version

As previously, one obtains a faster yet slightly less accurate algorithm by avoiding the computation of \( e_4 \). This gives Algorithm 12 below.

Algorithm 12 – \( 3\text{Prod}_{2,3}^{\text{fast}}(x_0, x_1, y_0, y_1, y_2) \). (37 operations & 1 test)

Require: \( \bar{x} \) DW and \( \bar{y} \) TW ; \( p \geq 6 \)
Ensure: \( \bar{z} \) TW and \( |\bar{r} - \bar{z}\bar{y}| / |\bar{x}\bar{y}| \leq 18u^3 + 75u^4 \)

[same 5 first lines as Algorithm 9]

\[
\begin{align*}
z_{3,1} &\leftarrow \text{RN}(z_{10} + x_0 y_2) \quad (\text{FMA}) \\
z_3 &\leftarrow \text{RN}(z_{3,1} + z_{0,1}) \\
s_3 &\leftarrow \text{RN}(c, z_3) \\
e_0, e_1, e_2, e_3 &\leftarrow \text{VecSum}(z_{0,0}, b_0, b_1, s_3) \\
r_0 &\leftarrow e_0 \\
r_1, r_2 &\leftarrow \text{VSEB}(2)(e_1, e_2, e_3) \\
\text{return} \; (r_0, r_1, r_2)
\end{align*}
\]

Similarly to what was done before, we obtain, provided that \( p \geq 6 \),

\[
|\bar{r} - \bar{z}\bar{y}| / |\bar{x}\bar{y}| \leq 18u^3 + 75u^4.
\]

This bound is very tight: in binary64 arithmetic, with the input values given by (3), we obtain a relative error around \( (18 - 2.4 \cdot 10^{-6})u^3 \).

8 Reciprocal of a TW number

To compute the reciprocal of a TW, we use Algorithm 13 below, which is based on the Newton-Raphson iteration, following the idea of [8] for general expansions. This algorithm requires the stronger constraint \( p \geq 10 \). For computing \( 1/x \), the Newton-Raphson iteration is

\[
r_{n+1} = r_n(2 - r_n x)
\]

which ensures a quadratic convergence towards \( 1/x \) as soon as \( r_0 \) is close enough to \( 1/x \).

Algorithm 13 – \( 3\text{Reci}(x_0, x_1, x_2) \). (73 operations & 2 tests and relative error bounded by \( 11.5u^3 + 1465u^4 \) if \( 3\text{Prod}_{2,3} \) is Algorithm 11, 65 operations & 1 test and relative error bounded by \( 19u^3 + 1502u^4 \) if \( 3\text{Prod}_{2,3} \) is Algorithm 12.

Require: \( \bar{x} \) TW ; \( p \geq 10 \)

\[
a \leftarrow \text{RN}((1 + 2u)/x_0) \\
h_{1,1} &\leftarrow 2\text{Prod}_2(1 + 2u)(a, x_0) = \text{RN}(ax_0 - (1 + 2u)) \quad (\text{FMA}) \\
h_1 &\leftarrow \text{RN}(-h_{1,1} - ax_1) \quad (\text{FMA}) \\
b_{0,1}, b_{1,1} &\leftarrow 2\text{Prod}(a, 1 - 2u) \\
b_{1,2} &\leftarrow \text{RN}(b_{1,1} + ah_1) \quad (\text{FMA}) \\
b &\leftarrow \text{Fast2Sum}(b_{0,1}, b_{1,2}) \\
n &\leftarrow 2 - 3\text{Prod}_{2,3}(b, \bar{x}) \\
\bar{y} &\leftarrow 3\text{Prod}_{2,3}(b, n) \\
\text{return} \; (y_0, y_1, y_2)
\]

A natural starting point for the Newton-Raphson iterations would be \( \text{RN}(1/x_0) \), obtained through a floating-point division. However, Algorithm 13 uses \( \text{RN}((1 + 2u)/x_0) \) instead of \( \text{RN}(1/x_0) \) to start the calculations in order to take profit from the fact that for any FP number \( x \), \( \text{RN}(x \times \text{RN}(1/x_0)) = 1 + 2u \) (the proof is straightforward).

Thus in Algorithm 13, one has to imagine a “virtual” \( h_0 = 2 - (1 + 2u) = 1 - 2u \). This defines a DW number \( \bar{h} = (h_0, h_1) \) that approximates \( 2 - a \cdot (x_0 + x_1) \).

Several variants are possible for implementing the last two lines of Algorithm 13. First, one can choose the accurate (Algorithm 11) or fast (Algorithm 12) version of \( 3\text{Prod}_{2,3} \) to implement the DW×TW products. Then, further simplification is possible: to compute \( 2 - 3\text{Prod}_{2,3}(\bar{b}, \bar{x}) \) there is no need to actually perform a multiplication followed by a subtraction, one can just slightly modify the multiplication algorithm, essentially by replacing the terms \( e_i \) in Algorithm 11 or Algorithm 12 by their opposite by turning some \( + \) into \( - \) operations and conversely in order not to waste any operation (this gives Algorithm 16 in the appendix, see supplementary material). The last product can also be simplified by taking into account the fact that \( i_0 = 1 \). We have,

Theorem 9. If \( \bar{x} \) is a TW, and if \( p \geq 10 \), then the relative error committed by \( 3\text{Reci}(\bar{x}) \) is bounded by \( 11.5u^3 + 1465u^4 \) if \( 3\text{Prod}_{2,3} \) is Algorithm 11 (accurate version), and by \( 19u^3 + 1502u^4 \) if \( 3\text{Prod}_{2,3} \) is Algorithm 12 (fast version).

The proof of Theorem 9, with the modified versions of Algorithms 11 and 12 is given in the appendix.

9 Quotient of two TW numbers

To compute \( \bar{z}/\bar{x} \), the first idea that springs in mind is to compute the reciprocal of \( \bar{x} \) and then to multiply it by \( \bar{z} \). However (using the notation \( \bar{b} \) that appears in Algorithm 13), this would mean that we compute something like \( \bar{z} \times (\bar{b} \times (2 - \bar{b} \bar{x})) \), while it is significantly better to compute \( (\bar{z} \times \bar{b}) \times (2 - \bar{b} \bar{x}) \). In particular, this allows to parallelize the computations of \( \bar{z} \bar{b} \) and \( 2 - \bar{b} \bar{x} \). We obtain Algorithm 14 below.
Algorithm 14 – 3Div(z₀, z₁, z₂, x₀, x₁, x₂). (119 operations & 4 tests and relative error bounded by 24u³ + 1509u⁴ if the “accurate” multiplications algorithms are used, and 103 operations & 2 tests and relative error bounded by 39u³ + 1582u⁴ if the “fast” multiplications algorithms are used.)

Require: z, x TW ; p ≥ 10
[same 5 first lines as Algorithm 13]
\( \bar{b} \leftarrow \text{Fast2Sum}(b₀, b₁) \)
\( \hat{i} \leftarrow 2 - 3\text{Prod}_{2,3}(\bar{b}, \bar{x}) \)
\( \tilde{a} \leftarrow 3\text{Prod}_{2,3}(\tilde{b}, \tilde{x}) \)
\( \tilde{y} \leftarrow 3\text{Prod}_{3,3}(\tilde{a}, \hat{i}) \)
return \((y₀, y₁, y₂)\)

As for Algorithm 13, several variants are possible for implementing the products 3Prod₂,₃ and 3Prod₃,₃ in Algorithm 14. One can choose the accurate or fast version of the DW×TW and TW×TW multiplication algorithms. Again, the subtraction that appears in the line “ \( \hat{i} \leftarrow 2 - 3\text{Prod}_{2,3}(\bar{b}, \bar{x}) \)” does not need to be performed: the multiplication algorithm is slightly modified instead (Algorithm 16 in the appendix, see supplementary material).

For the computation of the final product, one of course can use Algorithm 9 or Algorithm 10. However, as for the reciprocal operation, these algorithms can be simplified significantly by taking into account that \( i₀ = 1 \). The corresponding multiplication algorithm in the “fast” case is Algorithm 18 in the appendix. We finally have,

Theorem 10. If \( x, \bar{x} \) are TW numbers and if \( p ≥ 10 \), then the relative error committed by 3Div(\( z, \bar{z} \)) is bounded by 24u³ + 1509u⁴ if the “accurate” versions of the multiplication algorithms are used, and by 39u³ + 1582u⁴ if the “fast” versions are used.

The proof is given in the appendix, see supplementary material.

10 \textit{Square root of a TW number}

To compute the square root of a TW number, we use Algorithm 15 below, based again on the Newton-Raphson iteration. Its analysis and the various possible optimizing tricks are very similar to the ones used before for division. The underlying iteration is now

\[
r_{n+1} = rₙ \left( \frac{3}{2} - \frac{1}{2} rₙx \right),
\]

which ensures a quadratic convergence towards \( 1/\sqrt{x} \) as soon as \( r₀ \) is close enough to \( 1/\sqrt{x} \).

Algorithm 15 – 3SqRt(x₀, x₁, x₂). (127 operations & 4 tests and and relative error bounded by 24u³ + 10260u⁴ if the “accurate” multiplications algorithms are used, and 111 operations & 2 tests and relative error bounded by 39u³ + 1033u⁴ if the “fast” multiplications algorithms are used.)

Require: x TW ; p ≥ 11
Ensure: \( y \) TW and \( |\frac{\sqrt{x}}{3} - y| ≤ 24u³ + 10260u⁴ \), resp. 39u³ + 1033u⁴
\( a \leftarrow \text{RN}\left((1 + 4u)/\text{RN}(\sqrt{x})\right) \)
\( a' = \frac{1}{2}a \) (exact)
\( h₀, h₁, h₂ \leftarrow 2\text{Prod}(a, x₀) \)
\( h₀^{(1)} \leftarrow \text{RN}(h₁ + a₁) \) (FMA)
\( h₀^{(2)} \leftarrow 2\text{Prod}(a', h₀^{(1)}) \)
\( h₀^{(2)} \leftarrow 1.5 - h₀^{(2)} \) (exact)
\( h₁^{(2)} \leftarrow -\text{RN}(h₁ + a₁h₀^{(1)}) \) (FMA)
\( b₀, b₁, b₂ \leftarrow 2\text{Prod}(b₀, h₀^{(2)}) \)
\( b₁, b₂ \leftarrow \text{RN}(b₁ + a₁b₂) \) (FMA)
\( \bar{b} \leftarrow \text{Fast2Sum}(b₀, b₁) \)
\( \bar{b}' = \frac{1}{2}\bar{b} \) (exact)
\( i₀^{(1)} \leftarrow 3\text{Prod}_{2,3}(\bar{b}, \bar{x}) \)
\( i₀^{(2)} \leftarrow 1.5 - 3\text{Prod}_{2,3}(\bar{b}', i₀^{(1)}) \)
\( \tilde{y} \leftarrow 3\text{Prod}_{3,3}(i₀^{(1)}, i₀^{(2)}) \)
return \((y₀, y₁, y₂)\)

We have,

Theorem 11. If \( x \) is a TW, then the relative error committed by 3SqRt\( \text{acc}(x) \) (resp. 3SqRt\( \text{fast}(x) \)) is bounded by 24u³ + 10260u⁴ (resp. 39u³ + 1033u⁴).

The major steps of the proof are given in the appendix, see supplementary material.

11 \textit{Implementation and Tests}

11.1 Experimental settings

We have implemented the multiplication and division algorithms presented in this paper (Algorithms 9 and 10 for multiplication, and Algorithm 14 in the “fast” and “accurate” versions for division). Our goal was twofold: to compare the performances of our algorithms with other solutions, and to assess the correctness and tightness of our error bounds. Our algorithms have been implemented in the Campary library [25]. We compare them to algorithms based on “general” floating-point expansions used in the special case \( n = 3 \), also implemented in the Campary library, and to the GNU MPFR [5] multiple-precision library. MPFR also provides reference values. The various considered algorithms are given in Table 1.

Source code of this benchmark is available by request to the authors. It has been compiled with GCC 8.2.1 20181127 with compilation options “-O3 -march=native”, libraries GMP 6.1.2, and MPFR 4.0.2 from Archlinux repositories were used. The precision
We checked the various considered TW and FP expansion where URandom has an uniform density probability in (3Div

Table 2, the “accurate” version of the division algorithm variability. In Table 2, we reported the average and values, their accuracies will not differ significantly.

11.2 Correctness and performance

The input operands were generated using the MPFR URandom algorithm. More precisely, each FP component of a TW operand \( x = (x_0, x_1, x_2) \) was obtained as:

\[
x_{i+1} = \text{URandom()} \times \text{ulp}(x_i),
\]

where URandom has an uniform density probability in the range \([0, 1]\).

### Table 1

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
<th>Known error bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>mpfr_mul</td>
<td>[25, Alg 31]</td>
<td>8u^3 + 48u^4 [25, Thm 3.4.4.]</td>
</tr>
<tr>
<td>baileyMul</td>
<td>[25, Alg 33]</td>
<td>8u^3 + 49u^4 [25, Thm 3.4.5.]</td>
</tr>
<tr>
<td>3Prod\textsuperscript{acc}</td>
<td>Alg 9</td>
<td>28u^3 + 107u^4 Thm 7</td>
</tr>
<tr>
<td>3Prod\textsuperscript{fast}</td>
<td>Alg 10</td>
<td>44u^3 + 176u^4</td>
</tr>
<tr>
<td>division</td>
<td>[25, Alg 36]</td>
<td></td>
</tr>
<tr>
<td>3Div\textsuperscript{acc}</td>
<td>Alg 14</td>
<td>24u^3 + 1509u^4 Thm 10</td>
</tr>
<tr>
<td>3Div\textsuperscript{fast}</td>
<td>Alg 14</td>
<td>39u^3 + 1582u^4 Thm 10</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Name</th>
<th>Relative error</th>
<th>Ratio to bound</th>
<th>Speed (Mop/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mpfr_mul</td>
<td>n/a</td>
<td></td>
<td>37.0 0.9</td>
</tr>
<tr>
<td>truncatedMul</td>
<td>3.42 \times 10^{-49}</td>
<td>0.031</td>
<td>48.3 0.9</td>
</tr>
<tr>
<td>baileyMul</td>
<td>3.42 \times 10^{-49}</td>
<td>0.031</td>
<td>64.5 1.2</td>
</tr>
<tr>
<td>3Prod\textsuperscript{acc}</td>
<td>2.12 \times 10^{-47}</td>
<td>0.554</td>
<td>66.4 1.2</td>
</tr>
<tr>
<td>3Prod\textsuperscript{fast}</td>
<td>2.69 \times 10^{-47}</td>
<td>0.446</td>
<td>80.3 1.7</td>
</tr>
<tr>
<td>mpfr_div</td>
<td>n/a</td>
<td></td>
<td>10.4 0.3</td>
</tr>
<tr>
<td>division</td>
<td>6.43 \times 10^{-49}</td>
<td>9.3</td>
<td>9.3 0.2</td>
</tr>
<tr>
<td>3Div\textsuperscript{acc}</td>
<td>1.56 \times 10^{-47}</td>
<td>0.476</td>
<td>13.2 0.3</td>
</tr>
<tr>
<td>3Div\textsuperscript{fast}</td>
<td>1.31 \times 10^{-47}</td>
<td>0.245</td>
<td>16.8 0.3</td>
</tr>
</tbody>
</table>

Hence, compared to standard \(n\)-word floating-point expansion algorithms with \(n = 3\), our triple-word algorithms offer faster arithmetic, at the price of a slight (yet carefully bounded) loss in accuracy. This is of interest for implementing correctly-rounded transcendental functions in floating-point arithmetic, since in general this requires a bit more than twice the target precision (i.e., double-word arithmetic is not enough, and very accurate triple-word arithmetic is an overkill).

### Conclusion

We have shown that usual floating-point expansion algorithms adapted for building and adding triple-word numbers are correct in the context of triple-words, and we have introduced algorithms for rounding, multiplying, reciprocating, dividing and computing square roots of triple-word numbers, along with their correctness proofs and error bounds.

Some of the error bounds have been shown to be tight (at least in binary64 arithmetic). Our algorithms have been implemented in the Campary library. Our experiments show that the obtained algorithms are faster than generic \(n\)-word arithmetics at the price of a slight (yet carefully bounded) loss in accuracy.

### Acknowledgement

We are grateful to the anonymous reviewers, whose comments have been very helpful for revising the original manuscript.

### References


