Arithmetic algorithms for extended precision using floating-point expansions

Mioara Joldes, Olivier Marty, Jean-Michel Muller, Valentina Popescu

To cite this version:
Mioara Joldes, Olivier Marty, Jean-Michel Muller, Valentina Popescu. Arithmetic algorithms for extended precision using floating-point expansions. IEEE Transactions on Computers, 2016, 65 (4), pp.1197 - 1210. 10.1109/TC.2015.2441714. hal-01111551v2

HAL Id: hal-01111551
https://hal.science/hal-01111551v2
Submitted on 2 Jun 2015

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.
Arithmetic algorithms for extended precision using floating-point expansions

Mioara Joldes¸, Olivier Marty, Jean-Michel Muller and Valentina Popescu

Abstract—Many numerical problems require a higher computing precision than the one offered by standard floating-point (FP) formats. One common way of extending the precision is to represent numbers in a multiple component format. By using the so-called floating-point expansions, real numbers are represented as the unevaluated sum of standard machine precision FP numbers. This representation offers the simplicity of using directly available, hardware implemented and highly optimized, FP operations. It is used by multiple-precision libraries such as Bailey’s QD or the analogue Graphics Processing Units (GPU) tuned version, GQD. In this article we briefly revisit algorithms for adding and multiplying FP expansions, then we introduce and prove new algorithms for normalizing, dividing and square rooting of FP expansions. The new method used for computing the reciprocal $a^{-1}$ and the square root $\sqrt{a}$ of a FP expansion $a$ is based on an adapted Newton-Raphson iteration where the intermediate calculations are done using “truncated” operations (additions, multiplications) involving FP expansions. We give here a thorough error analysis showing that it allows very accurate computations. More precisely, after $q$ iterations, the computed FP expansion $x = x_0 + \ldots + x_{q-1}$ satisfies, for the reciprocal algorithm, the relative error bound: $|(x - a^{-1})/a^{-1}| \leq 2^{-2^{q}(p-3)-1}$ and, respectively, for the square root one: $|x - 1/\sqrt{a}| \leq 2^{-2^{q}(p-3)-1}/\sqrt{a}$, where $p > 2$ is the precision of the FP representation used ($p = 24$ for single precision and $p = 53$ for double precision).

Index Terms—floating-point arithmetic, floating-point expansions, high precision arithmetic, multiple-precision arithmetic, division, reciprocal, square root, Newton-Raphson iteration

1 INTRODUCTION

Many numerical problems in dynamical systems or planetary orbit dynamics, such as the long-term stability of the solar system [1], finding sinks in the Henon Map [2], iterating the Lorenz attractor [3], etc., require higher precisions than the standard double precision (now called binary64 [4]). Quad or higher precision is rarely implemented in hardware, and the most common solution is to use software emulated higher precision libraries, also called arbitrary precision libraries. There are mainly two ways of representing numbers in higher precision. The first one is the multiple-digit representation: numbers are represented by a sequence of possibly high-radix digits coupled with a single exponent. An example is the representation used in GNU MPFR [5], an open-source C library, which, besides arbitrary precision, also provides correct rounding for each atomic operation. The second way is the multiple-term representation in which a number is expressed as the unevaluated sum of several standard floating-point (FP) numbers. This sum is usually called a FP expansion. Bailey’s library QD [6] uses this approach and supports double-double (DD) and quad-double (QD) computations, i.e. numbers are represented as the unevaluated sum of 2 or 4 standard double-precision FP numbers. The DD and QD formats and the operations implemented in that library are not compliant with the IEEE 754-2008 standard, and do not provide correctly rounded operations. However, this multiple-term representation offers the simplicity of using directly available and highly optimized hardware implemented FP operations. This makes most multiple-term algorithms straightforwardly portable to highly parallel architectures, such as GPUs. In consequence, there is a demand for algorithms for arithmetic operations with FP expansions, that are sufficiently simple yet efficient, and for which effective error bounds and thorough proofs are given. Several algorithms already exist for addition and multiplication [7], [6, Thm. 44, Chap. 14].

In this article we mainly focus on division (and hence, reciprocal) and square root, which are less studied in literature. For these algorithms we provide a thorough error analysis and effective error bounds. There are two classes of algorithms for performing division and square root: the so-called digit-recurrence algorithms [8], that generalize the paper-and-pencil method, and the algorithms based on the Newton-Raphson iteration [9], [10]. While the algorithms suggested so far for dividing expansions belong to the former class, here we will be interested in studying the possible use of the latter class:
since its very fast, quadratic convergence is appealing when high precision is at stake.

Another contribution of this article is a new method for the renormalization of FP expansions. This operation ensures certain precision related requirements and is an important basic brick in most computations with FP expansions. Our renormalization procedure takes advantage of the computer’s pipeline, so it is fast in practice. For the sake of completeness, we also briefly present a variant of addition and multiplication algorithms which we implemented, and for which we intend on providing a full error analysis in a future related article.

A preliminary version of our work concerning only the case of division was recently presented in [11].

The outline of the paper is the following: in Section 2 we recall some basic notions about FP expansions and the algorithms used for handling them. Then, in Section 3 we give the new renormalization algorithm along with the proof of correctness. In Section 4 we present methods for performing divisions, including existing algorithms based on long classical division on expansions (Sec. 4.1) and the Newton based method (Sec. 4.2), followed by the correctness proof, the error analysis and the complexity analysis. After that, in Section 5 we give a similar method for computing the square root of an expansion along with the complexity analysis of the algorithm. Finally, in Section 6 we assess the performance of our algorithms – in terms of number of FP operations and proven accuracy bounds.

## 2 Floating-point expansions

A normal binary precision-$p$ floating-point (FP) number is a number of the form

\[ x = M_x \cdot 2^{e_x + p + 1}, \]

with \(2^{p-1} \leq |M_x| \leq 2^p - 1\), where \(M_x\) is an integer. The integer \(e_x\) is called the exponent of \(x\), and \(M_x \cdot 2^{p+1}\) is called the *significand* of \(x\). We denote accordingly to Goldberg’s definition: ulp(x) = \(2^{e_x + p + 1}\) [7, Chap. 2] (ulp is an acronym for *unit in the last place*). Another useful concept is that of *ulp in the last significant place*: uls(x) = ulp(x) \(
\cdot 2^{z_x}\), where \(z_x\) is the number of trailing zeros at the end of \(M_x\).

In order to ensure the uniqueness of the representation we need to set the first bit of the significand to 1 and adjust the exponent according to that. This is called a *normalized* representation. This is not possible if \(x\) is less than \(2^{e_{\text{min}}}\), where \(e_{\text{min}}\) is the smallest allowed exponent. Such numbers are called *subnormal*, where the first bit of the significand is 0 and the exponent is the minimum representable one. The IEEE 754 standard specifies that an *underflow* exception is raised every time a subnormal number occurs and the operation is inexact.

A natural extension of the notion of DD or QD is the notion of *floating-point expansion*. The arithmetic on FP expansions was first developed by Priest [12], and in a slightly different way by Shewchuk [13]. If, starting from a set of FP inputs, we only perform exact operations, then the values we obtain are always equal to finite sums of FP numbers. Such finite sums are called expansions. A natural idea is to try to manipulate such expansions, for performing calculations that are either exact, either approximate yet very accurate.

### Definition 2.1
A FP expansion \(u\) with \(n\) terms is the unevaleduated sum of \(n\) FP numbers \(u_0, \ldots, u_{n-1}\), in which all nonzero terms are ordered by magnitude (i.e., \(u_i \neq 0 \Rightarrow |u_i| \geq |u_{i+1}|\)). Each \(u_i\) is called a *component* of \(u\).

The notion of expansion is “redundant” since a nonzero number always has more than one representation as a FP expansion. To make the concept useful in practice and easy to manipulate, we must introduce a constraint on the components: the \(u_i\)’s cannot “overlap”. The notion of overlapping varies depending on the authors. We give here two very different definitions, using the above-introduced notation.

### Definition 2.2
(Nonoverlapping FP numbers) Assuming \(x\) and \(y\) are normal numbers with representations \(M_x \cdot 2^{e_x + p + 1}\) and \(M_y \cdot 2^{e_y + p + 1}\) (with \(2^{p-1} \leq |M_x|, |M_y| \leq 2^p - 1\)), they are \(P\)-nonoverlapping (that is, nonoverlapping according to Priest’s definition [14]) if \(|e_y - e_x| \geq p\).

### Definition 2.3
An expansion is \(P\)-nonoverlapping (that is, nonoverlapping according to Priest’s definition [14]) if all of its components are mutually \(P\)-nonoverlapping.

A visual representation of Definition 2.3, inspired from [17] can be seen in Fig. 1 (a). Shewchuk [13] weakens this into *nonzero-overlapping* sequences as shown in Fig. 1 (b) (also inspired from [17]):

### Definition 2.4
An expansion \(u_0, u_1, \ldots, u_{n-1}\) is \(S\)-nonoverlapping (that is, nonoverlapping according to Shewchuk’s definition [13]) if for all \(0 < i < n\), we have \(e_{u_{i-1}} - e_{u_i} \geq p - z_{u_{i-1}}\).

![Fig. 1: Nonoverlapping sequence by (a) Priest’s scheme and (b) Shewchuk’s scheme [17].](image_url)
done in faithful FP arithmetic (see [14]), since round-to-nearest rounding mode was not so common. More recently, a slightly stronger sense of nonoverlapping was introduced in 2001 by Hida, Li and Bailey [6]:

**Definition 2.5.** An expansion $u_0, u_1, \ldots, u_{n-1}$ is $\mathcal{B}$-nonoverlapping (that is, nonoverlapping according to Bailey’s definition [6]) if for all $0 < i < n$, we have $|u_i| \leq \frac{1}{2} \text{ulp}(u_{i-1})$.

**Remark 2.6.** For $\mathcal{P}$-nonoverlapping expansions we have $|u_i| \leq (2^p - 1)/2^p \text{ulp}(u_{i-1})$ and for $\mathcal{S}$-nonoverlapping expansions $|u_i| \leq (2^p - 1)/2^p \text{uls}(u_{i-1})$.

Even though we presented here three different types of nonoverlapping, in what follows we will focus only on the $\mathcal{P}$ and $\mathcal{B}$-nonoverlapping expansions, since, in general, they provide more precision per given number of terms of a FP expansion.

### 2.1 Error free transforms

Most algorithms performing arithmetic operations on expansions are based on the so-called *error-free transforms* (EFT), that make it possible to compute both the result and the rounding error of a FP addition or multiplication. This implies that in general, each such EFT, applied to two FP numbers, still returns two FP numbers. Although these algorithms use native precision operations only, they keep track of all accumulated rounding errors, ensuring that no information is lost.

We present here two EFTs that we use as basic bricks for our work. Algorithm 2Sum (Algorithm 1) computes the exact sum of two FP numbers $a$ and $b$ and returns the result under the form $s + e$, where $s$ is the result rounded to nearest and $e$ is the rounding error. It requires only 6 native FP operations (flops), which it was proven to be optimal in [15], if we have no information on the ordering of $a$ and $b$.

**Algorithm 1 2Sum $(a, b)$.

\[
\begin{align*}
\pi & \leftarrow \text{RN}(a \cdot b) \\
// \text{RN stands for performing the operation in rounding to nearest mode.} \\
t & \leftarrow \text{RN}(s - b) \\
e & \leftarrow \text{RN}(\text{RN}(a - t) + \text{RN}(b - \text{RN}(s - t))) \\
\text{return } (s, e) \text{ such that } s = \text{RN}(a + b) \text{ and } s + e = a + b
\end{align*}
\]

There exists a similar EFT, that performs the same addition using only 3 native precision FP operations. This one is called Fast2Sum [7] and it requires the exponent of $a$ to be larger than or equal to that of $b$. This condition might be difficult to check, but of course, if $|a| \geq |b|$, it will be satisfied.

For multiplying two FP numbers there exist two algorithms: Dekker’s product and 2MultFMA. They compute the product of two FP numbers $a$ and $b$ and return the exact result as $\pi$, the result rounded to nearest plus $e$, the rounding error. The first one requires 17 flops. The most expensive part of the algorithm is the computation of the error $e = a \cdot b - \pi$, but if a *fused-multiply-add* (FMA [7]) instruction, that takes only one flop, is available, it is easily computed. This gives Algorithm 2MultFMA (Algorithm 2), that takes only 2 flops. This algorithm works providing that the product $a \cdot b$ does not overflow and $e_a + e_b \geq e_{\text{min}} + p - 1$, where $e_a$ and $e_b$ are the exponents of $a$ and $b$ and $e_{\text{min}}$ is the minimum representable exponent. If the second condition is not satisfied, the product may not be representable as the exact sum of two FP numbers ($e$ would be below the underflow threshold).

**Algorithm 2 2MultFMA $(a, b)$.

\[
\begin{align*}
\pi & \leftarrow \text{RN}(a \cdot b) \\
// \text{RN stands for performing the operation in rounding to nearest mode.} \\
e & \leftarrow \text{fma}(a, b, -\pi) \\
\text{return } (\pi, e) \text{ such that } \pi = \text{RN}(a \cdot b) \text{ and } \pi + e = a \cdot b
\end{align*}
\]

These EFT can be extended to work on several inputs by chaining, resulting in the so-called distillation algorithms [16]. From these we make use of an algorithm called VecSum by Ogita, Rump, and Oishi [13], [17]. VecSum, presented in Fig. 2 and Algorithm 3, is simply a chain of 2Sum that performs an EFT on $n$ FP numbers.

**Algorithm 3 VecSum $(x_0, \ldots, x_{n-1})$.

**Input:** $x_0, \ldots, x_{n-1}$ FP numbers.

**Output:** $e_0 + \cdots + e_{n-1} = x_0 + \cdots + x_{n-1}$.

\[
\begin{align*}
s_{n-1} & \leftarrow x_{n-1} \\
\text{for } i & \leftarrow n - 2 \text{ to } 0 \text{ do} \\
&s_i, e_{i+1} \leftarrow 2\text{Sum}(x_i, s_{i+1}) \\
\text{end for} \\
e_0 & \leftarrow s_0 \\
\text{return } e_0, \ldots, e_{n-1}
\end{align*}
\]

**Fig. 2:** VecSum with $n$ terms. Each 2Sum box performs Algorithm 1, the sum is outputted to the left and the error downwards.

### 2.2 Addition and multiplication algorithms for expansions

An algorithm that performs the addition of two expansions $a$ and $b$ with $n$ and $m$ terms, respectively, will return a FP expansion with at most $n + m$ terms. Similarly, for multiplication, the product can have at most $2nm$ terms [12]. So-called normalization algorithms are used to render the result nonoverlapping, and this also implies a potential reduction in the number of terms.

Many variants of algorithms that compute the sum and the product of two FP expansions have been presented in the literature [12], [13], [6], [16]. Here, we only...
briefly present the algorithms that we used in our actual implementation. The addition is based on the merge algorithm and the multiplication is a generalization of Bailey’s algorithm for DD and QD [6] and it was first presented in [2].

The addition presented in Algorithm 4 and Fig. 3 is performed by merging the two FP expansions, \(a\) with \(n\) and \(b\) with \(m\) terms, respectively, and normalizing the resultant array for obtaining an approximation \(s\) on \(r\) terms of the sum \(a + b\).

**Algorithm 4** Algorithm of addition of FP expansions.

**Input:** FP expansions \(a = a_0 + \ldots + a_{n-1}; b = b_0 + \ldots + b_{m-1}; r\) length of the result.

**Output:** FP expansion \(s = s_0 + \ldots + s_{r-1}\).

1: \(f[0 : m + n - 1] \leftarrow \text{Merge}(a[0 : n - 1], b[0 : m - 1])\)
2: \(s[0 : r - 1] \leftarrow \text{Renorm}(f[0 : m + n - 1], r)\)
3: return FP expansion \(s = s_0 + \ldots + s_{r-1}\).

Fig. 3: Addition of FP expansions with \(n\) and \(m\) terms. The Merge box performs a classic algorithm for merging two arrays and the Renormalize box performs Algorithm 6.

Subtraction is performed simply by negating the FP terms in \(b\).

**Algorithm 5** Algorithm of multiplication of FP expansions with \(k\) terms.

**Input:** FP expansions \(a = a_0 + \ldots + a_{k-1}; b = b_0 + \ldots + b_{k-1}\).

**Output:** FP expansion \(r = r_0 + \ldots + r_{k-1}\).

1: \((r_0, e_0) \leftarrow 2\text{MultFMA}(a_0, b_0)\)
2: for \(n \leftarrow 1\) to \(k - 1\) do
3:   for \(i \leftarrow 0\) to \(n\) do
4:     \((p_i, e_i) \leftarrow 2\text{MultFMA}(a_i, b_{n-i})\)
5:   end for
6:   \(r_n, e[0 : n^2 + n - 1] \leftarrow \text{VecSum}(p[0 : n], e[0 : n^2 - 1])\)
7:   \(e[0 : (n + 1)^2 - 1] \leftarrow e[0 : n^2 + n - 1], e[0 : n]\)
8: end for
9: for \(i \leftarrow 1\) to \(k - 1\) do
10:   \(r_k \leftarrow r_k + a_i \cdot b_{k-i}\)
11: end for
12: for \(i \leftarrow 0\) to \(k^2 - 1\) do
13:   \(r_k \leftarrow r_k + e_i\)
14: end for
15: \(r[0 : k - 1] \leftarrow \text{Renormalize}(r[0 : k])\)
16: return FP expansion \(r = r_0 + \ldots + r_{k-1}\).

In Fig. 4 and Algorithm 5 we present the multiplication algorithm. Although we have implemented fully customized versions, for simplicity, we give here only the “\(k\) input - \(k\) output” variant of the algorithm. We consider two expansions \(a\) and \(b\) each with \(k\) terms and we compute the \(k\) most significant components of the product \(r = a \cdot b\). In the renormalization step (line 15 of Algorithm 5), we use an extra error correction term, so we perform our “error free transformation scheme” \(k+1\) times.

Here, we just give an intuitive explanation of the multiplication algorithm. Let \(\varepsilon = \frac{1}{2} \text{ulp}(r_0)\). Roughly speaking, if \(r_0\) is of the order of \(\Lambda\), then \(e_0\) is of order \(\mathcal{O}(\varepsilon \Lambda)\). So for the product \((p, e) = 2\text{MultFMA}(a_i, b_j)\), \(p\) is of order \(\mathcal{O}(\varepsilon n \Lambda)\) and \(e\) is of order \(\mathcal{O}(\varepsilon n+1 \Lambda)\), where \(n = i + j\), and we consider only the terms for which \(0 \leq n \leq k\). This implies that for each \(n\) we have \(n + 1\) products to compute (line 4 of Algorithm 5). Next, we need to add all terms of the same order of magnitude. Beside the \(n+1\) products, we also have \(n^2\) terms resulting from the previous iteration. This addition is performed using VecSum (Algorithm 3) to obtain \(r_n\) in line 6. The remaining terms are concatenated with the errors from the \(n+1\) products, and the entire \(e_0, \ldots, e_{(n+1)^2-1}\) array is used in the next iteration. The \((k+1)\)-st component \(r_k\) is obtained by simple summation of all remaining errors with the simple products of order \(\mathcal{O}(\varepsilon^k \Lambda)\). EFT are not needed in the last step since the errors are not reused.

For the addition and multiplication algorithms presented in this section, we will provide an effective error analysis in a subsequent paper. An important step for this goal is to provide a thorough proof for the renormalization, which is used at the end of each of these two algorithms. So, in what follows we focus on our new algorithm for renormalization of expansions.

### 3 Renormalization Algorithm for Expansions

While several renormalization algorithms have been proposed in literature, Priest [12] algorithm seems to be the only one provided with a complete correctness proof. It has many conditional branches, which make it slow in practice, and has a worst case FP operation count of: \(R(n) = 20(n-1)\), for an input FP expansion with \(n\)-terms.

In an attempt to overcome the branching problem we developed a new algorithm (Algorithm 6), for which we provide a full correctness proof.

First, we need to define the concept of FP numbers that overlap by at most \(d\) digits.

**Definition 3.1.** Consider an array of FP numbers: \(x_0, x_1, \ldots, x_{n-1}\). According to Priest’s [12] definition, they overlap by at most \(d\) digits \((0 \leq d < p)\) if and only if \(\forall i, 0 \leq i \leq n - 2, \exists k_i, \delta_i\) such that:

\[
2^{k_i} \leq |x_i| < 2^{k_i+1}, \tag{1}
\]

\[
2^{k_i-d_i} \leq |x_{i+1}| \leq 2^{k_i-d_i+1}, \tag{2}
\]

\[
\delta_i \geq p - d, \tag{3}
\]

\[
\delta_i + \delta_{i+1} \geq p - z_{i-1}, \tag{4}
\]

where \(z_{i-1}\) is the number of trailing zeros at the end of \(x_{i-1}\) and for \(i = 0, z_{-1} := 0\).
prove several intermediate properties. The notations
simpler layers based on VecSum
trated in Fig. 5) is based on different layers of chained
x
Consider an array
P
algorithm returns a
m
numbers that overlap by at most
d
be an input parameter, with
m
1
2Sum
Algorithm 6 returns a "truncation" to
m
steps of a
FP expansion
Algorithm 6, illus-
Fig. 4: Multiplication of FP expansions with
k
terms. Each 2MultFMA box performs Algorithm 2, the product is outputted
downwards and the error to the right; the VecSum box performs Algorithm 3, in which the most significant component of the
sum is outputted downwards; the circled + and * signs represent standard round-to-nearest FP addition and multiplication.

Proposition 3.2. Let
x_0, x_1, \ldots, x_{n-1}
be an array of FP numbers which overlap by at most
d
digits (0 ≤ d < p).
The following properties hold:

\[ |x_{j+1}| \leq 2^d \ulp(x_j), \]
\[ \ulp(x_{j+1}) \leq 2^{d-p} \ulp(x_j), \]
\[ |x_{j+2} + x_{j+1}| \leq (2^d + 2^{2d-p}) \ulp(x_j). \]

Proof: We have \ulp(x_j) = 2^{k_j-p+1} and from (3) we get
\[ |x_{j+1}| < 2^{k_j-d_j+1} \ulp(x_j) < 2^d \ulp(x_j). \]
This proves that (5) holds for all 0 ≤ j < n - 1.

By applying (3) we get \ulp(x_{j+1}) = 2^{k_j-d_j-p+1} \ulp(x_j), which proves that (6) holds for all 0 ≤ j < n - 1.

We have
\[ |x_{j+1}| \leq 2^d \ulp(x_j) \]
and
\[ |x_{j+2}| \leq 2^{ulp(x_{j+1})} \]
from which (7) follows.

The renormalization algorithm (Algorithm 6, illustrated in Fig. 5) is based on different layers of chained 2Sum. For the sake of simplicity, these are grouped in simpler layers based on VecSum. We will prove that our algorithm returns a \( P \)-nonoverlapping sequence.

Proposition 3.3. Consider an array
x_0, x_1, \ldots, x_{n-1}
of FP numbers that overlap by at most
d ≤ p - 2 digits and let
m
be an input parameter, with 1 ≤ m ≤ n - 1. Provided that no underflow / overflow occurs during the calculations, Algorithm 6 returns a "truncation" to \( m \) terms of a \( P \)-nonoverlapping FP expansion \( f = f_0 + \ldots + f_{m-1} \) such that
\[ x_0 + \ldots + x_{n-1} = f. \]

To prove this proposition, in what follows, we first prove several intermediate properties. The notations
\[ a_0 b_0 \]

Algorithm 6 Renormalization algorithm

Input: FP expansion \( x = x_0 + \ldots + x_{n-1} \) consisting of FP numbers that overlap by at most \( d \) digits, with \( d ≤ p - 2 \); \( m \) length of output FP expansion.

Output: FP expansion \( f = f_0 + \ldots + f_{m-1} \) with \( f_{i+1} \leq \left( \frac{1}{2} + 2^{-p+2} + 2^{-p} \right) \ulp(f_i), \) for all \( 0 \leq i < m - 1. \)

1: \( [0 : n - 1] \leftarrow VecSum(x[0 : n - 1]) \)
2: \( f^{(0)}[0 : m] \leftarrow VecSumErrBranch(e[0 : n - 1], m + 1) \)
3: for \( i \leftarrow 0 \) to \( m - 2 \) do
4: \( f^{(i+1)}[i : m] \leftarrow VecSumErr(f^{(i)}[i : m]) \)
5: end for
6: return FP expansion \( f = f^{(1)}_0 + \ldots + f^{(m-1)}_{m-1}. \)

used in the proof \((s_i, e_i, \varepsilon_i, f_i, \rho_i, g_i)\) are defined on the schematic drawings of the algorithms discussed. We also raise the important remark that at each step we prove that all the 2Sum blocks can be replaced by Fast2Sum ones, but for simplicity of the proof we chose to present first the 2Sum version.

First level (line 1, Algorithm 6)
It consists in applying Algorithm 3, VecSum (see also Fig. 2) on the input array, from where we obtain the array \( e = (e_0, e_1, \ldots, e_{n-1}). \)

Proposition 3.4. After applying the VecSum algorithm, the output array \( e = (e_0, e_1, \ldots, e_{n-1}) \) is \( S \)-nonoverlapping and may contain interleaving zeros.

Proof: Since \( s_i = RN(x_i + s_{i+1}), s_i \) is closer to \( x_i + s_{i+1} \)
we deduce that From (8) and (9) we have: in [18] we get: and by using a property given by Jeannerod and Rump

**Remark**

We know that \( s_{j+1} = \text{RN}(x_{j+1} + \text{RN}(... + x_{n-1})) \) and by using a property given by Jeannerod and Rump in [18] we get:

\[
|s_{j+1} - (x_{j+1} + ... + x_{n-1})| 
\leq (n-j-2) \cdot 2^{-p} \cdot (|x_{j+1}| + ... + |x_{n-1}|). \tag{9}
\]

From (8) and (9) we have:

\[
|s_{j+1}| \leq 2^d \left( \frac{2^p}{2^p-1} \right) (1 + (n-j-2)2^{-p}) \text{ulp}(x_j).
\]

It is easily seen that

\[
2^d \left( \frac{2^p}{2^p-1} \right) (1 + (n-j-2)2^{-p}) \leq 2^{p-1}, \tag{10}
\]

is satisfied for \( p \geq 4 \) and \( n \leq 16 \), for \( p \geq 5 \) and \( n \leq 32 \) and so on. This includes all practical cases, when \( d \leq p-2 \), so that \( \text{ulp}(s_{j+1}) < \text{ulp}(x_j) \). Therefore \( x_j \) and \( s_{j+1} \) are multiples of \( \text{ulp}(s_{j+1}) \), thus \( x_j + s_{j+1} \) is multiple of \( \text{ulp}(s_{j+1}) \), hence \( \text{RN}(x_j + s_{j+1}) \) is multiple of \( \text{ulp}(s_{j+1}) \), and \( |e_{j+1}| = |x_j + s_{j+1} - \text{RN}(x_j + s_{j+1})| \) is multiple of \( \text{ulp}(s_{j+1}) \).

Also, by definition of 2Sum, we have \( |e_{j+2}| \leq 2 |e_{j+1}| \). Now, we can compare \( |e_{j+1}| \) and \( |e_{j+2}| \). Since \( |e_{j+1}| \) is a multiple of \( \text{ulp}(s_{j+1}) \), either \( e_{j+1} = 0 \) or \( e_{j+1} > 2 |e_{j+2}| \) and multiple of \( 2^k \), such that \( 2^k > |e_{j+2}| \). This implies that the array \( e = (e_0, e_1, ..., e_{n-1}) \) is S-nonoverlapping and may have interleaving zeros. □

**Remark 3.5.** Since we have \( |s_{j+1}| \leq 2^{p-1} \text{ulp}(x_j) \), for \( d \leq p-2 \) and \( p \geq 4 \) for \( n \) up to 16 and also \( \text{ulp}(x_j) \leq 2^{1-p} |x_j| \) we deduce that \( |s_{j+1}| \leq |x_j| \). Hence, at this level we can use instead of 2Sum basic blocks the Fast2Sum ones.

**Second level (line 2, Algorithm 6)**

**Algorithm 7** Second level of the renormalization algorithm - VecSumErrBranch

**Input:** S-nonoverlapping FP expansion \( e = e_0 + ... + e_{n-1} \); \( m \) length of the output expansion.

**Output:** FP expansion \( f = f_0 + ... + f_{m-1} \) with \( f_{j+1} \leq \text{ulp}(f_j), \ 0 \leq j < m-1 \).

1. \( j \leftarrow 0 \)
2. \( z_0 = e_0 \)
3. for \( i \leftarrow 0 \) to \( n-2 \) do
4. \( (f_{j}, e_{j+1}) \leftarrow \text{2Sum}(e_i, e_{i+1}) \)
5. if \( e_{j+1} \neq 0 \) then
6. \( j \leftarrow j + 1 \)
7. return FP expansion \( f = f_0 + ... + f_{m-1} \).

It is applied on the array \( e \) obtained previously. This is also a chain of 2Sum, but instead of starting from the least significant, we start from the most significant component. Also, instead of propagating the sums we propagate the errors. If however, the error after a 2Sum block is zero, then we propagate the sum (this is shown in Figure 6). In what follows we will refer to this algorithm by VecSumErrBranch (see Algorithm 7). The following property holds:

**Proposition 3.6.** Let an input array \( e = (e_0, ..., e_{n-1}) \) of S-non-overlapping terms and \( 1 \leq m \leq n \) the required number of output terms. After applying VecSumErrBranch, the output array of \( f = (f_0, ..., f_{m-1}) \), with \( 0 \leq m \leq n-1 \) satisfies \( |f_i| \leq \text{ulp}(f_i) \) for all \( 0 \leq i < m-1 \).

**Proof:** The case when \( e \) contains 1 or 2 elements is trivial. Consider now at least 3 elements. By definition of 2Sum, we have \( |e_1| \leq \frac{1}{2} \text{ulp}(f_0) \) and by definition of S-nonoverlapping,

\[
e_0 = E_0 \cdot 2^{k_0} \text{ with } |e_1| < 2^{k_0},
\]

\[
e_1 = E_1 \cdot 2^{k_1} \text{ with } |e_2| < 2^{k_1}.
\]

Hence, \( f_0 \) and \( e_1 \) are both multiples of \( 2^{k_1} \). Two possible cases may occur:

(i) \( e_1 = 0 \). If we choose to propagate directly \( e_1 = 0 \), then \( f_1 = e_2 \) and \( e_2 = 0 \). This implies by induction that \( f_i = e_{i+1}, \forall i \geq 1 \). So, directly propagating the error poses a problem, since the whole remaining chain of 2Sum is executed without any change. So, as shown in

Fig. 5: Renormalization of FP expansions with n terms. The VecSum box performs Algorithm 3, the VecSum-ErrBranch box, Algorithm 7 and the VecSumErr box, Algorithm 8.
Algorithm 7, line 11, when $\varepsilon_{i+1} = 0$ we propagate the sum $f_j$.

(ii) $\varepsilon_i \neq 0$. Then $|e_2| < |e_1|$ and $|e_1 + e_2| < 2|e_1|$, from where we get $|g_1| = |\text{RN}(e_1 + e_2)| \leq 2|e_1| \leq \text{ulp}(f_0)$.

We prove by induction the following statement: at step $i > 0$ of the loop in Algorithm 7, both $f_j - 1$ and $\varepsilon_i$ are multiples of $2^{k_i}$ with $|\varepsilon_{i+1}| < 2^{k_i}$. We proved above that $i = 1$ holds. Suppose now it holds for $i$ and prove it for $i + 1$. Since $f_j - 1$ and $\varepsilon_i$ are multiples of $2^{k_i}$ with $|\varepsilon_{i+1}| < 2^{k_i}$ and $e_{i+1} = E_{i+1} \cdot 2^{k_{i+1}}$ with $|e_{i+2}| < 2^{k_{i+1}}$ (by definition of $S$-nonoverlapping), it follows that both $f_j$ and $\varepsilon_{i+1}$ are multiples of $2^{k_{i+1}}$ (by definition of 2Sum).

Finally, we prove the relation between $f_j$ and $f_j - 1$. If $\varepsilon_{i+1} = 0$, we propagate $f_j$, i.e. $\varepsilon_i = f_j - 1$. Otherwise $|\varepsilon_{i+1}| < |\varepsilon_i|$, so $|\varepsilon_{i+1} + \varepsilon_i| < 2|\varepsilon_i|$ and finally $|f_j| = |\text{RN}(e_{i+1} + \varepsilon_i)| \leq 2|\varepsilon_i| \leq \text{ulp}(f_j - 1)$.

Remark 3.7. For this algorithm we can also use Fast2Sum instead of 2Sum. We already showed that either $|\varepsilon_{i+1}| < |\varepsilon_i|$, or $\varepsilon_i = 0$, in which case we replace $\varepsilon_i = f_j - 1$, which is a multiple of $2^{k_i}$ with $|\varepsilon_{i+1}| < 2^{k_i}$.

Third level and further (lines 3-5, Algorithm 6)
On the previously obtained array we apply a similar chain of 2Sum, starting from the most significant component and propagating the error. In these subsequent levels, no conditional branching is needed anymore (see Algorithm 8).

Algorithm 8 Third level of the renormalization algorithm - VecSumErr

Input: FP expansion $f = f_0 + \ldots + f_m$ with $|f_{i+1}| \leq \text{ulp}(f_i)$, for all $0 \leq i \leq m - 1$.

Output: FP expansion $g = g_0 + \ldots + g_m$ with $|g_i| \leq \left(\frac{1}{2} + 2^{-p+2}\right)\text{ulp}(g_0)$ and $|g_{i+1}| \leq \text{ulp}(g_i)$, for $0 < i \leq m - 1$.

1: $\rho_0 = f_0$
2: for $i \leftarrow 0$ to $m - 1$ do
3: \hspace{1cm} $(g_i, \rho_{i+1}) \leftarrow 2\text{Sum}(\rho_i, f_{i+1})$
4: end for
5: $g_m \leftarrow \varepsilon_m$
6: return FP expansion $g = g_0 + \ldots + g_m$.

We prove the following property:

Proposition 3.8. After applying Algorithm 8, VecSumErr on an input array $f = (f_0, \ldots, f_m)$, with $|f_{i+1}| \leq \text{ulp}(f_i)$, for all $0 \leq i \leq m - 1$, the output array $g = (g_0, \ldots, g_m)$ satisfies $|g_1| \leq \left(\frac{1}{2} + 2^{-p+2}\right)\text{ulp}(g_0)$ and $|g_{i+1}| \leq \text{ulp}(g_i)$, for $0 < i \leq m - 1$.

Proof: Since $|f_1| \leq \text{ulp}(f_0)$ and $g_0 = \text{RN}(f_0 + f_1)$ we have: $(1/2)\text{ulp}(f_0) \leq \text{ulp}(g_0) \leq 2\text{ulp}(f_0)$, and $|\rho_1| \leq (1/2)\text{ulp}(g_0)$. We also have:

$$|f_1| \leq \text{ulp}(f_0),$$

which implies $\text{ulp}(f_1) \leq 2^{-p+1}\text{ulp}(f_0)$,

$$|f_2| \leq \text{ulp}(f_1) \leq 2^{-p+2}\text{ulp}(g_0).$$

Hence:

$$|\rho_1 + f_2| \leq (1/2 + 2^{-p+2})\text{ulp}(g_0).$$

Since $(1/2 + 2^{-p+2})\text{ulp}(g_0)$ is a FP number, we also have:

$$|g_1| = |\text{RN}(\rho_1 + f_2)| \leq (1/2 + 2^{-p+2})\text{ulp}(g_0).$$

This bound is very close to $(1/2)\text{ulp}(g_0)$ and it seems that in most practical cases, one actually has $|g_1| \leq 1/2\text{ulp}(g_0)$. This implies that $g_0$ and $g_1$ are “almost” $B$-nonoverlapping and a simple computation shows that they are $B$-nonoverlapping as soon as $p \geq 3$, which occurs in all practical cases. As we iterate further, we get:

$$|\rho_{i+1}| \leq (1/2)\text{ulp}(g_i),$$

$$|f_{i+1}| \leq \text{ulp}(f_i),$$

which implies $\text{ulp}(f_{i+1}) \leq 2^{-p+1}\text{ulp}(f_i)$.

We know that $\rho_i$ is a multiple of $\text{ulp}(f_i)$ and from this we derive two cases: (i) $\rho_0 = 0$, and as a consequence $\forall j \geq i, g_j = f_{j+1}$ and $g_m = 0$. In the second case (ii) we get:

$$|f_{i+1}| \leq |\rho_i| \leq (1/2)\text{ulp}(g_{i-1}),$$

$$|f_{i+1} + \rho_i| \leq \text{ulp}(g_{i-1}),$$

$$|g_i| = |\text{RN}(f_{i+1} + \rho_i)| \leq \text{ulp}(g_{i-1}).$$

Remark 3.9. For Algorithm 8 we can also use the faster algorithm, Fast2Sum$(\rho_i, f_{i+1})$, because we either have $\rho_i = 0$ or $|f_{i+1}| \leq |\rho_i|$. The above proposition shows that while we obtain a nonoverlapping condition for the first two elements of the resulting array $g$, for the others we don’t strengthen the existing bound $|g_{i+1}| \leq \text{ulp}(g_i)$. There is an advantage however: if zeros appear in the summation process, they are pushed at the end; we don’t use any branching. This suggests to continue applying a subsequent level of the same algorithm on the remaining elements, say $g_1, \ldots, g_m$. This is the idea of applying $m - 1$ levels of VecSumErr in lines 3-5, Algorithm 6. We are now able to prove Prop. 3.3.

Proof: (of Prop. 3.3) Consider $m \geq 2$, otherwise the output reduces to only one term. The loop in lines 3-5 of
Algorithm 6 is executed at least once. From Prop. 3.4, 3.6 and 3.8 we deduce that $|f_1^{(1)}| \leq (1/2 + 2^{-p+2}) \ulp(f_0^{(1)})$ and $|f_{i+1}^{(1)}| \leq \ulp(f_i^{(1)})$, for $i > 0$. If $m = 2$ then $f_0^{(1)}$, $f_1^{(1)}$ are outputted and the proposition is proven. Otherwise, $f_0^{(1)}$ is kept unchanged and another VecSumErr is applied to remaining $f_1^{(1)}, \ldots, f_m^{(1)}$. We have:

$$
\begin{align*}
|f_1^{(2)}| & \leq (1/2 + 2^{-p+2}) \ulp(f_0^{(1)}), \\
|f_2^{(2)}| & \leq \ulp(f_1^{(2)}) \leq 2^{-p+1} (1/2 + 2^{-p+2}) \ulp(f_0^{(1)}), \\
|f_3^{(2)}| & \leq 2^{-p+1} \ulp(f_0^{(1)}).
\end{align*}
$$

Hence,

$$
|f_1^{(2)}| = \left| \text{RN}(f_1^{(1)} + f_2^{(1)}) \right|, \\
\leq (1/2 + 2^{-p+2} + 2^{-p+1}) \ulp(f_0^{(1)}).
$$

Similarly,

$$
\begin{align*}
|f_2^{(2)}| & \leq (1/2 + 2^{-p+2}) \ulp(f_2^{(2)}), \\
|f_3^{(2)}| & \leq \ulp(f_2^{(2)}) \leq 2^{-p+1} (1/2 + 2^{-p+2}) \ulp(f_1^{(2)}), \\
& \leq 2^{-p+1} \ulp(f_0^{(1)}).
\end{align*}
$$

So, $f_0^{(1)}, f_1^{(2)}, f_2^{(2)}$ are nonoverlapping. It follows by induction that after $m - 1$ loop iterations the output $f_0^{(1)}, \ldots, f_{m-2}^{(1)}, f_{m-1}^{(1)}$ is a $P$-nonoverlapping expansion. Finally, when all $n - 1$ terms are considered, after at most $n - 1$ loop iterations we have: $x_0 + \ldots + x_{n-1} = f_0^{(1)} + \ldots + f_{n-2}^{(1)} + f_{n-1}^{(1)}$.

Fig. 8 gives an intuitive drawing showing the different constraints between the FP numbers before and after the first two levels of Algorithm 6. The notation is the same as in Fig. 5.

Table 1 shows some effective values of the worst case FP operation count for Priest’s renormalization algorithm [12] and Algorithm 6. One can see that for $n \leq 7$ our algorithm performs better or the same. Even though from values of $n > 7$ Algorithm 6 performs worse in terms of operation count than Priest’s one, in practice, the last $m - 1$ levels will take advantage of the computers pipeline, because we do not need branching conditions anymore, which makes it faster in practice.

In what follows we denote by $AddRoundE(x[0 : n - 1], y[0 : m - 1], r)$, an algorithm for expansions addition, which given two $(P- or B-)$ nonoverlapping expansions, returns the $r$ most significant terms of the exact normalized $(P- or B-)$ nonoverlapping sum. If no request is made on the number of terms to be returned, then we denote simply by $AddE(x[0 : n - 1], y[0 : m - 1])$. Similarly, we denote by $MulRoundE, MulE, SubRoundE, SubE, DioRoundE, RenormalizeE$ algorithms for multiplication, subtraction, division and normalization.

### 4 Reciprocal Algorithm

#### 4.1 Algorithms using classical long division on expansions

In reference [12], division is done using the classical long division algorithm (a variation of the paper-and-pencil method), which is recalled in Algorithm 9.

**Algorithm 9** Priest’s [12] division algorithm. We denote by $f[0 : \ldots]$ and expansion $f$ whose number of terms is not known in advance.

**Input:** FP expansion $a = a_0 + \ldots + a_{n-1}; b = b_0 + \ldots + b_{m-1};$ length of output quotient FP expansion $d$.

**Output:** FP expansion $q = q_0 + \ldots + q_d$ with at most $d$ terms s.t. $|q_n/b| < 2^{-(n-d)/p}.$

1: $q_0 = \text{RN}(a_0/b_0)$
2: $r^{(0)}[0 : n - 1] \leftarrow a[0 : n - 1]$ 
3: for $i \leftarrow 1$ to $d - 1$ do
4: $f[0 : \ldots] \leftarrow \text{MulE}(q_{i-1}, b[0 : m - 1])$
5: $r^{(i)}[0 : \ldots] \leftarrow \text{RenormalizeE}(\text{SubE}(r^{(i-1)}[0 : \ldots], f[0 : \ldots]), f[0 : \ldots])$
6: $q_i = \text{RN}(r^{(i)})/b_0$
7: end for
8: $q[0 : \ldots] \leftarrow \text{RenormalizeE}(q[0 : d - 1])$
9: return FP expansion $q = q_0 + \ldots$
the end, on \( q_0 + q_1 + q_2 + \ldots \) in order to ensure non-overlapping. No error bound is given in [6].

Note that in Algorithm 9 [12] a renormalization step is performed after each computation of \( r = r - q_kb \). An error bound is given in [12]:

**Proposition 4.1.** [12] Consider two \( \mathcal{P} \)-nonoverlapping expansions: \( a = a_0 + \ldots + a_{n-1} \) and \( b = b_0 + \ldots + b_{m-1} \). Priest division algorithm [12] computes a quotient expansion \( q = q_0 + \ldots + q_{d-1} \) s.t.

\[
\frac{q - a/b}{a/b} < 2^{-d(p-1)} \prod_{i=0}^{d-1} (4i + 6).
\]

Daumas and Finot [19] modify Priest’s division algorithm by using only estimates of the most significant component of the remainder \( r_0 \) and storing the less significant components of the remainder and the terms \(-q_kb\) unchanged in a set that is managed with a priority queue. While the asymptotic complexity of this algorithm is better, in practical simple cases Priest’s algorithm is faster due to the control overhead of the priority queue [19]. The error bound obtained with Daumas’ algorithm is (using the same notations as above):

\[
\frac{q - a/b}{a/b} < 2^{-d(p-1)} \prod_{i=0}^{d-1} (4i + 6).
\]

### 4.2 Reciprocal of expansions with an adapted Newton-Raphson iteration

The classical Newton-Raphson (NR) iteration for computing reciprocals of \( \mathcal{P} \) expansions with truncated operations involving FP expansions, with truncated operations involving FP expansions. Our algorithm works with both \( \mathcal{B} \) - and \( \mathcal{P} \)-nonoverlapping FP expansions. For the sake of clarity we consider first the case of \( \mathcal{B} \)-nonoverlapping FP expansions, and then make the necessary adjustments for \( \mathcal{P} \)-nonoverlapping expansions in Proposition 4.4.

**Algorithm 10** Truncated Newton iteration based algorithm for reciprocal of an FP expansion.

**Input:** FP expansion \( a = a_0 + \ldots + a_{2^k-1} \); length of output FP expansion \( 2^q \).

**Output:** FP expansion \( x = x_0 + \ldots + x_{2^q-1} \) s.t. \(|x - 1/a| \leq 2^{-2^q(p-3)-1}/|a|\).

1. \( x_0 = \text{RN}(1/a_0) \)
2. for \( i \leftarrow 0 \) to \( q - 1 \) do
3. \( \hat{v}[0 : 2^{i+1} - 1] \leftarrow \text{MulRoundE}(x[0 : 2^i - 1], a[0 : 2^{i+1} - 1], 2^{i+1}) \)
4. \( \hat{v}[0 : 2^{i+1} - 1] \leftarrow \text{SubRoundE}(2, \hat{v}[0 : 2^{i+1} - 1], 2^{i+1}) \)
5. \( x[0 : 2^{i+1} - 1] \leftarrow \text{MulRoundE}(x[0 : 2^i - 1], \hat{v}[0 : 2^{i+1} - 1], 2^{i+1}) \)
6. end for
7. return FP expansion \( x = x_0 + \ldots + x_{2^q-1} \).

### 4.3 Error analysis of Algorithm 10

Let \( a = a_0 + \ldots + a_{2^k-1} \) be a \( \mathcal{B} \)-nonoverlapping FP expansion with \( 2^k \) terms and \( q \geq 0 \). We will prove that our algorithm returns an approximation \( x = x_0 + \ldots + x_{2^q-1} \) of \( 1/a \), in the form of a \( \mathcal{B} \)-nonoverlapping FP expansion with \( 2^q \) terms, such that

\[
|x - 1/a| \leq 2^{-2^q(p-3)-1}/|a|.
\]

We will first prove the following proposition:

**Proposition 4.2.** Consider a \( \mathcal{B} \)-nonoverlapping expansion \( u = u_0 + u_1 + \ldots + u_k \) with \( k \) normal binary FP terms of precision \( p \). Denote \( u(i) = u_0 + u_1 + \ldots + u_i, i \geq 0 \), i.e. “a truncation” of \( u \) to \( i+1 \) terms. The following inequalities hold for \( 0 \leq i \leq k \):

\[
|u_i| \leq 2^{-2^p} |u_0|,
\]

\[
|u - u(i)| \leq 2^{-2^p} |u| \frac{\eta}{1-\eta},
\]

\[
\left(1 - \frac{2^{-2^p} \eta}{1-\eta}\right) |u| \leq |u(i)| \leq \left(1 + \frac{2^{-2^p} \eta}{1-\eta}\right) |u|,
\]

\[
\frac{1}{|u - 1/u_0|} \leq \eta/|u|,
\]

where

\[
\eta = \sum_{j=0}^{\infty} 2^{(-j-1)p} = \frac{2^{-p}}{1-2^{-p}}.
\]

**Proof:** By definition of a \( \mathcal{B} \)-nonoverlapping expansion and since for any normal binary FP number \( u \), \( \text{ulp}(u) \leq 2^{-p+1} |u| \) we have \( |u| \leq (1/2) \text{ulp}(u_{i-1}) \leq 2^{-p} |u_{i-1}| \) and (16) follows by induction.

Consequently we have

\[
|u - u_0| = |u_1 + u_2 + \ldots + u_k| \leq 2^{-p} |u_0| + 2^{-2p} |u_0| + \ldots + 2^{-kp} |u_0| \leq |u_0| \eta.
\]

One observes that \( u \) and \( u_0 \) have the same sign. A possible proof is by noticing that \( 1 - \eta > 0 \) and \(-|u_0| \eta \leq u - u_0 \leq |u_0| \eta \). Suppose \( u_0 > 0 \), then

\[
-|u_0| \eta \leq u - u_0 \leq |u_0| \eta \quad \text{hence} \quad u_0(1 - \eta) \leq u \leq u_0(1 + \eta)
\]

which implies \( u > 0 \). The case \( u_0 < 0 \) is similar. It follows that

\[
|u|/(1 + \eta) \leq |u_0| \leq |u|/(1 - \eta).
\]
For (17) we use (20) together with:
\[
|u - u^{(i)}| \leq \sum_{j=0}^{\infty} 2^{-i-j-1}p |u_0| \leq 2^{-i\eta} |u_0|,
\]
and (18) is a simple consequence of (17). Similarly, (19) follows from \(|1/u - 1/u_0| = (1/|u|) \cdot |(u_0 - u)/u_0| \leq \eta/|u|\).

Proposition 4.3. Provided that no underflow / overflow occurs during the calculations, Algorithm 10 is correct when run with B-nonoverlapping expansions.

Proof: The input of the algorithm is a non-overlapping FP expansion \(a = a_0 + a_1 \cdots + a_{2^k-1}\), in which every term \(a_i\) is a normal binary FP number of precision \(p\). Let \(f_i = 2^{i+1} - 1\) and \(a(f_i) = a_0 + a_1 + \cdots + a_f\), i.e. “a truncation” of \(a\) to \(f_i + 1\) terms, with \(0 \leq i\).

For computing \(1/a\) we use Newton iteration: \(x_0 = \text{RN}(1/a_0)\), \(x_{i+1} = x_i(2 - a(f_i) x_i)\), \(i \geq 0\) by truncating each FP expansion operation as follows:

- let \(v_i := (a(f_i) \cdot x_i)\) be the exact product represented as a non-overlapping expansion on \(2(2i+1)\) terms, we compute \(\hat{v}_i := v_i^{(2i)}\) i.e. \(v_i\) “truncated to” \(2i+1\) terms;
- let \(w_i := 2 - \hat{v}_i\) be the exact result of the subtraction represented as a non-overlapping expansion on \(2i+1 + 1\) terms, we compute \(\hat{w}_i := w_i^{(2i)}\) i.e. \(v_i\) “truncated to” \(2i+1\) terms;
- let \(\tau_i := x_i \cdot \hat{w}_i\) be the exact product represented as a non-overlapping expansion on \(2 \cdot 2^{i+1}\) terms, we compute \(x_{i+1} := \tau_i^{(2i+1)}\) i.e. \(\tau_i\) “truncated to” \(2^{i+1}\) terms.

Let us first prove a simple upper bound for the approximation error in \(x_0\):

\[
\varepsilon_0 = |x_0 - 1/a| \leq 2\eta/|a|.
\]  
(21)

Since \(x_0 = \text{RN}(1/a_0)\), then \(|x_0 - 1/a_0| \leq 2^{-p} |1/a_0|\), so \(|x_0 - 1/a| \leq 2^{-p} |1/a_0| + |1/a - 1/a_0| \leq ((1 + \eta)2^{-p} + \eta)/|a| \leq 2\eta/|a|\) (from (20)).

Let us deduce an upper bound for the approximation error in \(x\) at step \(i + 1\), \(\varepsilon_{i+1} = |x_{i+1} - 1/a|\). For this, we will use a chain of triangular inequalities that make the transition from our “truncated” Newton error to the “untruncated” one. Let \(\gamma_i = 2^{-1-(2^{i+1}-1)p}\eta/(1 - \eta)\). We have from Proposition 4.2, eq. (17):

\[
|x_{i+1} - \tau_i| \leq \gamma_i |x_i \cdot \hat{w}_i|,
\]  
(22)

\[
|w_i - \hat{w}_i| \leq \gamma_i |w_i| \leq \gamma_i |2 - \hat{v}_i|,
\]  
(23)

\[
|v_i - \hat{v}_i| \leq \gamma_i |a(f_i) \cdot x_i|,
\]  
(24)

\[
|a - a(f_i)| \leq \gamma_i |a|.
\]  
(25)

From (22) we have:

\[
\varepsilon_{i+1} \leq |x_{i+1} - \tau_i| + |\tau_i - 1/a| \leq \gamma_i |x_i \cdot \hat{w}_i| + |x_i \cdot \hat{w}_i - 1/a|.
\]

\[
\gamma_i |x_i \cdot \hat{w}_i| + \gamma_i |x_i \cdot \hat{w}_i| + |x_i \cdot \hat{w}_i - 1/a| \leq (1 + \gamma_i) |x_i| |w_i - \hat{w}_i| + \gamma_i |x_i| w_i| + \gamma_i |x_i| w_i| + |x_i \cdot w_i - 1/a|.
\]

Using (23) and (24):

\[
\varepsilon_{i+1} \leq |x_i \cdot w_i - 1/a| + ((1 + \gamma_i) |x_i| w_i| + |x_i \cdot w_i - 1/a|.
\]

\[
\gamma_i |(1 + \gamma_i) |x_i| ((2 - v_i)| + |v_i - \hat{v}_i|) + \gamma_i |x_i| w_i| + |x_i \cdot w_i - 1/a|.
\]

\[
\gamma_i |x_i| (2 - a(f_i) \cdot x_i) - 1/a| + \gamma_i |(1 + \gamma_i) w_i| + \gamma_i |x_i| w_i| + |x_i \cdot w_i - 1/a|.
\]

By (25), we have:

\[
|x_i \cdot (2 - a(f_i) \cdot x_i) - 1/a| \leq |a| |x_i - 1/a| + \gamma_i |x_i| w_i| + |x_i \cdot w_i - 1/a|.
\]

\[
|x_i| |a(f_i)| \leq (1 + \gamma_i) |x_i| w_i| + |x_i \cdot w_i - 1/a|.
\]

Hence we have:

\[
\varepsilon_{i+1} \leq |x_i \cdot (2 - a(f_i) \cdot x_i) - 1/a| + \gamma_i |x_i| w_i| + |x_i \cdot w_i - 1/a| + 1/|a|.
\]

(26)

We now prove by induction that for all \(i \geq 0\):

\[
\varepsilon_i = |x_i - 1/a| \leq 2^{-2((p-1) - 1)}/|a|.
\]

(27)

For \(i = 0\), this holds from (21) and the fact that \(\eta = 1/(2^p - 1) \leq 2^{-p+1}\). For the inductive step, we have from (26):

\[
\varepsilon_{i+1} \leq (1 + \gamma_i) |a| |x_i - 1/a|^2 + \gamma_i (1 + \gamma_i)^2 (2 + \gamma_i) |a| |x_i|^2 |a| + \gamma_i (2 + \gamma_i) 1/|a|.
\]

(28)

which implies

\[
\frac{|a| \varepsilon_{i+1}}{2 - 2^{2(p+1)(p-3)}} \leq (1 + \gamma_i)^2 + (1 + 2^{1-p+2}) (2 + \gamma_i) \cdot (1 + 1 + \gamma_i)^2 \left(1 + 2^{2(p-3)-1}\right)^2 \leq 1/2.
\]

(29)

This completes our proof.
Proposition 4.4. Algorithm 10 is correct when run with \(\mathcal{P}\)-nonoverlapping expansions.

Proof: The previous analysis holds provided that we use Remark 2.6. This mainly implies the following changes \(\eta' = \frac{2}{2^p-\delta}, \gamma' = \left(\frac{2}{2^p-1}\right)^{2^p-1-1} \eta'\). With this change it is easy to verify that equations (21)–(28) hold as soon as \(p > 2\). Note that for the induction step at \(i = 1\), a tighter bound is needed for \(\varepsilon'_0 \leq \frac{2^p(1+\gamma'\eta')}{|a|} \leq \frac{2\eta'}{\eta' - 1} \), but the rest of the proof is identical, safe for some tedious computations.

4.4 Complexity analysis for reciprocal

Our algorithm has the feature of using “truncated” expansions, while some variants of Add-RoundE and Mul-RoundE compute the result fully and only then truncate. This is the case of Priest’s algorithms, which are not tuned for obtaining “truncated” expansions on the fly – and thus penalize our algorithm-\(\zeta\). On the other hand, the algorithms presented in Sec. 2.2, can take into account only the significant terms of the input expansions in order to compute the result. Even though these algorithms have not yet been proven to work properly, we obtained promising results in practice, so we will perform the complexity analysis based on them.

We present here the operation count of our algorithms, by taking ([7]) 6 FP operations for 2Sum, 3 for Fast2Sum and 2 for 2MultFMA. For the sake of simplicity, for multiplication, we will consider that the input expansions have the same number of terms.

- The renormalization (Algorithm 6) of an overlapping expansion \(x\) with \(n\) terms, requires \((2n-3)+\sum_{i=0}^{n-1} m - i\) Fast2Sum calls and \(n-1\) comparisons. This accounts for \(R_{new}(n, m) = 7n + \frac{3}{2}m^2 + \frac{3}{2}m - 13\) FP operations.
- The addition (Algorithm 4) of two \(\mathcal{P}\)-nonoverlapping expansions requires \(n + m - 1\) comparisons and a renormalization \(R_{new}(n + m, r)\). This accounts for \(A(n, m, r) = \frac{3}{2}r^2 + \frac{3}{2}r + 8n + 8m - 14\) FP operations.
- The multiplication (Algorithm 5) of two \(\mathcal{P}\)-nonoverlapping expansions requires \(\sum_{i=1}^{k} 2\text{MultFMA}\) calls, \(\sum_{i=0}^{k-1} (n^2 + n) \) 2Sum calls, \(k - 1\) FP multiplications, followed by \(k^2 + k - 2\) FP additions and a renormalization \(R_{new}(k + 1, k)\) in the end. This accounts for \(M(k) = 2k^3 + \frac{3}{2}k^2 + \frac{19}{2}k - 9\) FP operations.

The special case of multiplying a FP expansion to a single FP value accounts for only \(M1(k) = \frac{3}{2}k^2 + \frac{13}{2}k - 7\).

Using these algorithms for addition and multiplication of FP expansions, Priest’s division (Algorithm 9) requires \(d\) divisions and \((d - 1)(M1(k)) + \sum_{i=0}^{d-1} A(k + 2k(i - 1), k + 2k(i - 1), k + 2k(i - 1)) + R_{new}(d, d)\) function calls in the worst case. This accounts for \(D(d, k) = 2d^3k^2 - 6dk^2 + \frac{3}{2}d^2k + \frac{9}{2}dk^2 + 8d > 3d^2k^2 - \frac{9}{2}dk^2 - \frac{9}{2}dk + \frac{1}{2}k - 6\) FP operations.

Proposition 4.5. Using for addition and multiplication of FP expansions the algorithms presented in Sec. 2.2, Algorithm 10 requires \(\frac{3}{2}s^q + \frac{3}{4}s^q + 57 \cdot 2^q - 24q - \frac{1510}{27}\) FP operations.

Proof: During the \(i\)th iteration the following operations with expansions are performed: \(2\) multiplications \(M(2^{i+1}); \) one addition \(A(2^{i+1}, 1, 2^{i+1})\). Since \(q\) iterations are done, the total number of FP operations is: \(\frac{3}{2}s^q + \frac{3}{4}s^q + 57 \cdot 2^q - 24q - \frac{1510}{27}\) FP operations.

Remark 4.6. Division is simply performed with Algorithm 10 followed by a multiplication \(M(2^q)\) where the numerator expansion has \(2^q\) terms.

5 Square root algorithms

The families of algorithms most commonly used are exactly the same as for division, although, in the case of FP expansions the digit-recurrence algorithm is typically more complicated than for division. This is why a software implementation would be tedious. Moreover, Newton-Raphson based algorithms offer the advantage of assuring a quadratic convergence.

5.1 Square root of expansions with an adapted Newton-Raphson iteration

Starting from the general Newton-Raphson iteration (13), we can compute the square root in two different ways. We can look for the zeros of the function \(f(x) = x^2 - a\) that leads to the so called “Heron iteration”:

\[x_{n+1} = \frac{1}{2}(x_n + \frac{a}{x_n})\]  

If \(x_0 > 0\), then \(x_n\) goes to \(\sqrt{a}\). This iteration needs a division at each step, which counts as a major drawback.

To avoid performing a division at each step we can look for the positive root of the function \(f(x) = 1/x^2 - a\). This gives the iteration

\[x_{n+1} = \frac{1}{2}(x_n + \frac{a}{x_n^2})\]  

This iteration converges to \(1/\sqrt{a}\), provided that \(x_0 \in (0, \sqrt{3}/\sqrt{a})\). The result can be multiplied by \(a\) in order to get an approximation of \(\sqrt{a}\). To obtain fast, quadratic, convergence, the first point \(x_0\) must be a close approximation to \(1/\sqrt{a}\). The division by 2 is done by multiplying each of the terms of the input expansion by 0.5, separately.

As in the case of the reciprocal (Sec. 4.2), in Algorithm 11 we use an adaption of iteration (31), using the truncated algorithms presented above.

The error analysis for this algorithm follows the same principle as the one for the reciprocal algorithm. The detailed proof is given in Appendix A. We show that the relative error decreases after every loop of the algorithm, by taking into account the truncations performed after each operation. The strategy is to make the exact Newton iteration term and bound appear. We show that by the end of the \(i\)th iteration of the loop, \(\varepsilon_i = \left|x_i(2^i) - 1/\sqrt{a}\right| \leq 2^{-2(p-3)} - 1/\sqrt{a}\).

In his library, QD, Bailey also uses the Newton iteration for the square root computation. Although he uses the same function as we do, he uses the iteration
Algorithm 11 Truncated “division-free” Newton iteration (31) based algorithm for reciprocal of the square root of an FP expansion. By “division-free” we mean that we do not need a division of FP expansions.

Input: FP expansion $a = a_0 + \ldots + a_{2^k-1}$; length of output FP expansion $2^q$.

Output: FP expansion $x = x_0 + \ldots + x_{2^n-1}$ s.t. $|x - 1/\sqrt{a}| \leq 2^{-2^q(p-3)-1}/\sqrt{a}$. (32)

1: $x_0 = \text{RN}(1/\sqrt{a_0})$
2: for $i \leftarrow 0$ to $q - 1$ do
3: $\tilde{v}[0 : 2^i+1 - 1] \leftarrow \text{MulRoundE}(x[0 : 2^i - 1], a[0 : 2^i+1 - 1], 2^i+1)$
4: $\tilde{w}[0 : 2^i+1 - 1] \leftarrow \text{MulRoundE}(x[0 : 2^i - 1], \tilde{v}[0 : 2^i+1 - 1], 2^i+1)$
5: $\tilde{y}[0 : 2^i+1 - 1] \leftarrow \text{SubRoundE}(3, \tilde{w}[0 : 2^i+1 - 1], 2^i+1)$
6: $\tilde{z}[0 : 2^i+1 - 1] \leftarrow \text{MulRoundE}(x[0 : 2^i - 1], \tilde{y}[0 : 2^i+1 - 1], 2^i+1)$
7: $x[0 : 2^i+1 - 1] \leftarrow x[0 : 2^i+1 - 1] \cdot 0.5$
8: end for
9: return FP expansion $x = x_0 + \ldots + x_{2^n-1}$.

under the form: $x_{i+1} = x_i + \frac{1}{2} x_i (1 - ax_i^2)$, which from a mathematical point of view is the same, but it requires a different implementation. Even though Bailey does not provide an error analysis for his algorithm, we managed to prove that the error bound is preserved when using this iteration (see Appendix A for the detailed proof).

“Heron iteration” algorithm

The same type of proof as above can be applied for the algorithm using the “Heron iteration” (30) and the same type of truncations. In this case (Algorithm 12) we obtain a slightly larger error bound for both types of nonoverlapping FP expansions: $|x - \sqrt{a}| \leq 3\sqrt{a} \cdot 2^{-2^q(p-3)-2}$. (33)

Algorithm 12 Truncated “Heron iteration” (30) based algorithm for square root of an FP expansion.

Input: FP expansion $a = a_0 + \ldots + a_{2^k-1}$; length of output FP expansion $2^q$.

Output: FP expansion $x = x_0 + \ldots + x_{2^n-1}$ s.t. $|x - \sqrt{a}| \leq 3\sqrt{a} \cdot 2^{-2^q(p-3)-2}$. (33)

1: $x_0 = \text{RN}(\sqrt{a_0})$
2: for $i \leftarrow 0$ to $q - 1$ do
3: $\tilde{v}[0 : 2^i+1 - 1] \leftarrow \text{DivRoundE}(a[0 : 2^i+1 - 1], x[0 : 2^i - 1], 2^i+1)$
4: $\tilde{w}[0 : 2^i+1 - 1] \leftarrow \text{AddRoundE}(x[0 : 2^i - 1], \tilde{v}[0 : 2^i+1 - 1], 2^i+1)$
5: $x[0 : 2^i+1 - 1] \leftarrow x[0 : 2^i+1 - 1] \cdot 0.5$
6: end for
7: return FP expansion $x = x_0 + \ldots + x_{2^n-1}$.

5.2 Complexity analysis for square root

We will perform our operation count based on the addition and multiplication presented in Sec. 2.2, the same as in Section 4.4.

Proposition 5.1. Using for addition, multiplication and division of FP expansions the algorithms previously presented, Algorithm 11 requires $\frac{48}{7} 8^q + 16 \cdot 4^q + 78 \cdot 2^q - 33q - \frac{699}{7} FP operations.

Proof: During the $i$th iteration, 3 multiplications $M(2^i+1)$, one addition $A(2^i+1, 1, 2^i+1)$ and one division by 2 are performed. Since $q$ iterations are done, the total number of FP operations is: $\frac{48}{7} 8^q + 16 \cdot 4^q + 78 \cdot 2^q - 33q - \frac{699}{7}.

Remark 5.2. We obtain the square root of an expansion by simply multiplying the result obtained from Algorithm 11 by the input expansion $a$. This means an additional $M(2^q)$, where $2^q$ is the number of terms in $a$.

Proposition 5.3. Using for addition, multiplication and division of FP expansions the algorithms previously presented, Algorithm 12 requires $\frac{368}{49} 8^q + \frac{196}{9} 4^q + 170 \cdot 2^q - 12q^2 - \frac{2245}{21} q \cdot \frac{87445}{441}$ FP operations.

Proof: One addition $A(2^i+1, 2^i+1, 2^i+1)$, one division $D(2^i+1)$ and a division by 2 are performed during each $i$th iteration. Since $q$ iterations are done, the total number of FP operations is: $\frac{368}{49} 8^q + \frac{196}{9} 4^q + 170 \cdot 2^q - 12q^2 - \frac{2245}{21} q \cdot \frac{87445}{441}$.

Based on these values, Algorithm 11 performs slightly better than Algorithm 12, and in the same time the obtained error bound is tighter.

6 COMPARISON AND DISCUSSION

Table 2 shows values of the bounds provided by our error analysis, compared with those of Priest and Daumas for the reciprocal computation. Our algorithm performs better for the same number of terms in the computed quotient, say $d = 2^q$ in equations (11) and (12). Moreover, our algorithm provides a unified error bound with quadratic convergence independent of using underlying $P$- or $B$-nonoverlapping expansions. In the last column of the same table we give the largest errors that we actually obtained through direct computation of the reciprocal using our algorithm. The given value represents the obtained value upper rounded to the immediate power of 2. For each table entry we performed 1 million random tests.

The complexity analysis shows that our algorithm performs better, for expansions with more than 2 terms, even if no error bound is requested (see Table 3 for some effective values of the worst case FP operation count).

Note that, for instance, to guarantee an error bound of $2^{-d(p-3)-1}$, Priest’s algorithm (based on the bound given in Prop 4.1) needs at least $(dp - 3d + 2)p/(p - 4)$ terms, which entails a very poor complexity. This implies that Daumas’ algorithm might be a good compromise in this
case, provided that the priority queue used there can be efficiently implemented.

This plus the performance tests that we ran confirm our hypothesis that for higher precisions the Newton-Raphson iteration is preferable to classical division.

In the case of the square root, because no error bound is given for the digit-recurrence algorithm we can only compare between the errors that we obtain if using the two different types of Newton iteration available for computing the square root. The effective values of the bounds are given in Table 4. The bound provided for Algorithm 11 is only slightly tighter that the one for Algorithm 12. The same as for the reciprocal, in the last column we present the bounds obtained through direct computation using Algorithm 11.

In Table 5 we give some effective values of the worst case FP operation count for Algorithm 11 vs Algorithm 12 based on section 5.2.

The algorithms presented in this article were implemented in the CAMPARY (CudA Multiple Preci-

### TABLE 2: Error bounds values for Priest (11) vs. Daumas (12) vs. our analysis (15). \( \beta \) gives the largest obtained errors for Algorithm 10 using the standard FP formats double and single. *underflow occurs.

<table>
<thead>
<tr>
<th>( p )</th>
<th>( q )</th>
<th>( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>53, 3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>53, 3</td>
<td>1</td>
<td>3.2</td>
</tr>
<tr>
<td>53, 3</td>
<td>2</td>
<td>3.2</td>
</tr>
<tr>
<td>53, 3</td>
<td>3</td>
<td>2.4</td>
</tr>
<tr>
<td>53, 3</td>
<td>4</td>
<td>2.8</td>
</tr>
<tr>
<td>24, 0</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>24, 0</td>
<td>1</td>
<td>3.2</td>
</tr>
<tr>
<td>24, 0</td>
<td>2</td>
<td>3.2</td>
</tr>
<tr>
<td>24, 0</td>
<td>3</td>
<td>3.2</td>
</tr>
<tr>
<td>24, 0</td>
<td>4</td>
<td>3.2</td>
</tr>
</tbody>
</table>

### TABLE 3: FP operation count for Priest vs. our algorithm; \( d = 2^8 \) terms are computed in the quotient.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>24</th>
<th>1714</th>
<th>52898</th>
<th>1808698</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alg. 9</td>
<td>22</td>
<td>1714</td>
<td>52898</td>
<td>1808698</td>
</tr>
<tr>
<td>Alg. 10 + Alg. 5</td>
<td>140</td>
<td>795</td>
<td>4693</td>
<td>31601</td>
</tr>
</tbody>
</table>

The algorithms presented in this article were implemented in the CAMPARY (CudA Multiple Preci-

### TABLE 4: Error bounds values for (44) vs. (33). \( \beta \) gives the largest obtained errors for Algorithm 11 using the standard FP formats double and single. *underflow occurs.

<table>
<thead>
<tr>
<th>( p )</th>
<th>( q )</th>
<th>( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>53, 3</td>
<td>0</td>
<td>2.93</td>
</tr>
<tr>
<td>53, 3</td>
<td>1</td>
<td>3.2</td>
</tr>
<tr>
<td>53, 3</td>
<td>2</td>
<td>2.93</td>
</tr>
<tr>
<td>53, 3</td>
<td>3</td>
<td>2.93</td>
</tr>
<tr>
<td>53, 3</td>
<td>4</td>
<td>2.93</td>
</tr>
<tr>
<td>24, 0</td>
<td></td>
<td>2.93</td>
</tr>
<tr>
<td>24, 0</td>
<td>1</td>
<td>3.2</td>
</tr>
<tr>
<td>24, 0</td>
<td>2</td>
<td>2.93</td>
</tr>
<tr>
<td>24, 0</td>
<td>3</td>
<td>2.93</td>
</tr>
<tr>
<td>24, 0</td>
<td>4</td>
<td>2.93</td>
</tr>
</tbody>
</table>

### TABLE 5: FP operation count for Algorithm 11 vs. Algorithm 12; \( d = 2^8 \) terms are computed in the quotient.

<table>
<thead>
<tr>
<th>( q )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alg. 11 + Alg. 5</td>
<td>182</td>
<td>1054</td>
<td>6275</td>
<td>42430</td>
</tr>
<tr>
<td>Alg. 12</td>
<td>170</td>
<td>1049</td>
<td>5972</td>
<td>39239</td>
</tr>
</tbody>
</table>

As a future work we intend to generalize the theoretical analysis of DD and QD addition/multiplication algorithms and thus to be able to provide a full error analysis for these algorithms.

### ACKNOWLEDGMENTS

The Authors would like to thank Région Rhône-Alpes and ANR FastRelax Project for the grants that support this activity.
References


Mioara Joldes was born in Abrud, Romania, in 1984. She received her Ph.D. degree in 2011 from the Ecole Normale Superieure de Lyon. She is Chargée de Recherches (junior researcher) at CNRS, France. Her research interests are in Computer Arithmetic, Validated Computing and Computer Algebra.

Olivier Marty was born in Evry, France, in 1993. He received his L3 degree in 2014 from the Ecole normale superieure de Cachan and he is now in the Parisian Master of Research in Computer Science at this school. His research interest is in Algorithm Design and Theoretical Computer science.

Jean-Michel Muller was born in Grenoble, France, in 1961. He received his Ph.D. degree in 1985 from the Institut National Polytechnique de Grenoble. He is Directeur de Recherches (senior researcher) at CNRS, France, and he is the co-head of GDR-IM. His research interests are in Computer Arithmetic. Dr. Muller was co-program chair of the 13th IEEE Symposium on Computer Arithmetic (Asilomar, USA, June 1997), general chair of SCAN'97 (Lyon, France, sept. 1997), general chair of the 14th IEEE Symposium on Computer Arithmetic (Adelaide, Australia, April 1999), general chair of the 22nd IEEE Symposium on Computer Arithmetic (Lyon, France, June 2015). He is the author of several books, including “Elementary Functions, Algorithms and Implementation” (2nd edition, Birkhäuser, 2006), and he coordinated the writing of the “Handbook of Floating-Point Arithmetic” (Birkhäuser, 2010). He is an associate editor of the IEEE Transactions on Computers, and a senior member of the IEEE.

Valentina Popescu was born in Baia-Mare, Romania, in 1991. She received her MS. degree in 2014 from the Universite Toulouse 3 - Paul Sabatier. She is a Ph.D. student in the AriC team, LIP Laboratory, ENS-Lyon, France. Her research interests are in Computer Arithmetic and Validated Computing.
**APPENDIX**

**Square root algorithm convergence proof**

This proof follows the same principle as the convergence proof for the reciprocal algorithm, and it uses the same notations. We will show that Algorithm 11, having as an input a FP expansion \( a = a_0 + \cdots + a_{2^k-1} \), computes and returns an approximation of \( \frac{1}{\sqrt{a}} \) in the form of a FP expansion \( x = x_0 + \cdots = x_{2^k-1} \), s.t.:

\[
\left| x - \frac{1}{\sqrt{a}} \right| \leq \frac{2^{-2^{(p-3)-1}}}{\sqrt{a}}.
\]

For the sake of clarity we present first the case of B-
nonoverlapping FP expansion, and then make the necessary adjustments for \( \mathcal{P} \)-nonoverlapping expansion in A.2.

**Proposition A.1.** Algorithm 11 is correct when run with B-
nonoverlapping expansions.

**Proof:** In Proposition 4.2 we gave and proved some properties of the B-
nonoverlapping FP expansions. To those we are going to add a new one:

\[
\left| \frac{1}{\sqrt{a}} - \frac{1}{\sqrt{u_0}} \right| \leq \frac{1}{\sqrt{a}} \eta.
\]

(34)

It can be seen that \( \frac{1}{\sqrt{a}} - \frac{1}{\sqrt{u_0}} = \frac{1}{\sqrt{a}} \left| 1 - \sqrt{\frac{u_0}{a}} \right| \). By using Proposition 4.2 eq. (20), the fact that \( u \) and \( u_0 \) have the same sign, and the fact that the square root is an increasing function, we have: \( 1 - \frac{\sqrt{u_0}}{\sqrt{a}} \leq 1 - \sqrt{\frac{1+\eta}{\eta}} \), which proves the property.

We continue by first proving a simple upper bound for the approximation error in \( x_0 \):

\[
\varepsilon_0 = \left| x_0 - \frac{1}{\sqrt{a}} \right| \leq \frac{1}{\sqrt{a}} \eta(3+\eta).
\]

(35)

We denote \( a = \text{RN}(\sqrt{a_0}) \), so we have \( x_0 = \text{RN}(1/\alpha) \). We know that \( \alpha - \sqrt{a_0} \leq 2^{-p} \sqrt{a_0} \) and \( |x_0 - \frac{1}{\alpha}| \leq 2^{-p} \frac{1}{\alpha} \), so we obtain:

\[
\left| \frac{1}{\alpha} - \frac{1}{\sqrt{a_0}} \right| \leq \left| \frac{1}{\alpha} - \frac{1}{\sqrt{a_0}} \right| \leq \frac{2^{-p} \frac{1}{\alpha}}{\sqrt{a_0}}.
\]

(36)

From (34) it follows:

\[
\varepsilon_0 \leq \left| x_0 - \frac{1}{\sqrt{a}} \right| \leq \frac{2 \eta}{\sqrt{a_0}}.
\]

Now we can conclude that \( \varepsilon_0 \leq (2\eta(1 + \frac{7}{2}) + \eta) \frac{1}{\sqrt{a}} \leq \eta(3 + \eta) \frac{1}{\sqrt{a}} \).

Before going further, let \( E_i = \varepsilon_i \sqrt{a} \), such that:

\[
E_0 \leq \eta(3 + \eta).
\]

Next we will deduce an upper bound for the approximation error in \( x \) at step \( i + 1 \), \( \varepsilon_{i+1} = \left| x_{i+1} - \frac{1}{\sqrt{a}} \right| \). For this, we use, same as in the case of the reciprocal, a chain of triangular inequalities that make the transition from our “truncated” Newton error, to the “untruncated” one.

For the truncations we use the same type of notations as for the reciprocal error bound proof, so: at each step of the algorithm \( v_0, v_1, \) and \( y_0, y_1 \) represent the exact results of the computations and we use the hatted notation \((\hat{v}_i, \hat{w}_i, \) and \( \hat{y}_i)\) to represent the results truncated to \( 2^{i+1} \) terms.

\( \tau_i \) denotes the exact result of the last two operations: \( \frac{1}{2} x_i \cdot \hat{y}_i \), so \( x_{i+1} \) is \( \tau_i \)’s truncation to \( 2^{i+1} \) terms.

Let \( \gamma_i = 2^{-2^{(i+1)-1}p} \frac{\eta}{\sqrt{a}} \), the same as before. We have from Proposition 4.2, eq. (17):

\[
\left| x_{i+1} - \tau_i \right| \leq \gamma_i \left| x_{i+1} \right| \leq \gamma_i \left| x_i \right| \left| \frac{1}{2} x_i \hat{y}_i \right|,
\]

(36)

\[
\left| y_i - \hat{y}_i \right| \leq \gamma_i \left| y_i \right| \left| \frac{3 - \hat{w}_i}{\hat{w}_i} \right|,
\]

(37)

\[
\left| w_i - \hat{w}_i \right| \leq \gamma_i \left| w_i \right| \left| \frac{1}{\hat{v}_i} \hat{w}_i \right|,
\]

(38)

\[
\left| \hat{v}_i - \frac{\hat{v}_i}{\hat{v}_i} \right| \leq \gamma_i \left| \hat{v}_i \right| \left| \frac{x_i a(f_i)}{x_i a(f_i)} \right|,
\]

(39)

\[
\left| a - a(f_i) \right| \leq \gamma_i |a|.
\]

(40)

From (36) we have:

\[
\varepsilon_{i+1} \leq \left| x_{i+1} - \tau_i \right| \leq \gamma_i \left| \frac{1}{2} x_i \hat{y}_i \right| + \frac{1}{2} x_i \hat{y}_i - \frac{1}{\sqrt{a}}.
\]

(41)

Using (37) and (38):

\[
\varepsilon_{i+1} \leq \left| \gamma_i(2 + \gamma_i) \frac{1}{2} x_i(3 - \hat{w}_i) \right| + \left| \frac{1}{2} x_i (3 - \hat{w}_i) - \frac{1}{\sqrt{a}} \right|
\]

\[
\leq \gamma_i(2 + \gamma_i) \frac{1}{2} x_i \left| (3 - w_i) + \gamma_i |w_i| \right| + \gamma_i \left| \frac{1}{2} x_i w_i \right| + \left| \frac{1}{2} x_i (3 - w_i) - \frac{1}{\sqrt{a}} \right|
\]

(42)

By (39) we have:

\[
\varepsilon_{i+1} \leq \left| \gamma_i \frac{1}{2} x_i \left| (2 + \gamma_i) |3 - x_i \hat{v}_i| \right| \right| + (1 + \gamma_i(2 + \gamma_i)) \left| \frac{1}{2} x_i (3 - \hat{v}_i) \right| + \left| \frac{1}{2} x_i (3 - \hat{v}_i) - \frac{1}{\sqrt{a}} \right|
\]

\[
\leq \gamma_i \left| \frac{1}{2} x_i \left| (2 + \gamma_i) |3 - x_i v_i| + \gamma_i |v_i x_i| \right| \right| + (1 + \gamma_i(2 + \gamma_i)(1 + \gamma_i)) \left| \frac{1}{2} x_i v_i \right| + \left| \frac{1}{2} x_i (3 - v_i) - \frac{1}{\sqrt{a}} \right|
\]

(43)

Because \( \sqrt{\frac{3}{\sqrt{a}}} \leq \sqrt{a_0}, \) then \( \frac{1}{\sqrt{a_0}} \leq \frac{1 + u}{\sqrt{a}} = \frac{1}{\sqrt{a}}. \)
From (40) we have:

$$
\varepsilon_{i+1} \leq \gamma_i (2 + \gamma_i) \left( |x_{i+1} - x_i| + \frac{1}{2} \frac{\gamma_i (3 - x_i^2)}{a} \right) + \gamma_i (\gamma_i + 1)^2 |x_{i+1} - x_i| + \gamma_i (\gamma_i + 1)^3 \frac{1}{2} x_i^2 + \gamma_i (\gamma_i + 1) \frac{1}{\sqrt{a}}.
$$

Hence we have:

$$
\varepsilon_{i+1} \leq (1 + \gamma_i)^2 \left| x_{i+1} - x_i \right| + \gamma_i (\gamma_i + 1)^2 + (\gamma_i + 1)^3 + \gamma_i (2 + \gamma_i) \frac{1}{\sqrt{a}}.
$$

By using the quadratic convergence of the sequence we can say that:

$$
\left| x_{i+1} - \frac{1}{\sqrt{a}} \right| = \frac{1}{2} \sqrt{a} \left| x_i \sqrt{a} + 2 \right| \left| x_i - \frac{1}{\sqrt{a}} \right|^2.
$$

We now prove by induction that for all \( i \geq 0 \) \( \varepsilon_i = \left| x_i - \frac{1}{\sqrt{a}} \right| \) respects the imposed bound.

We know that \( |x_i \sqrt{a}| \leq \varepsilon_i \sqrt{a} + 1 \) and \( |x_i^3 \sqrt{a}| \leq \frac{(\varepsilon_i \sqrt{a} + 1)^3}{\sqrt{a}} \) and from (41) we have:

$$
\varepsilon_{i+1} \leq \frac{1}{2} (1 + \gamma_i)^2 \varepsilon_i \left| x_{i+1} \right| + \frac{1}{2} \gamma_i ((\gamma_i + 1)^2 + (\gamma_i + 1)^3 + \gamma_i (2 + \gamma_i) \frac{1}{\sqrt{a}}.
$$

Using the notation \( E_i = \varepsilon_i \sqrt{a} \) we can transform (43) in an equation independent of \( a \):

$$
E_{i+1} \leq \frac{1}{2} (1 + \gamma_i)^2 (E_i + 3) E_i^2 + \gamma_i (\gamma_i + 1)^2 + (\gamma_i + 1)^3 + (\gamma_i + 1)^4) E_i + \gamma_i (2 + \gamma_i)
$$

For the last part of the proof we denote by \( f \) a function that writes the previous inequality as: \( E_{i+1} \leq f(E_i, i) \).

We want to show that \( \forall i \in \mathbb{N}, E_i \leq 2^{-2^p(p - 3) - 1} \) so we will define \( \text{ind}(i) = 2^{-2^p(p - 3) - 1} \).

For \( i = 0 \) we verify that \( E_0 \leq \text{ind}(0) \) for \( p \geq 3 \).

For \( i \geq 1 \) by induction:

- for \( i = 1 \) we can prove by using computer algebra that the inequality is verified if \( p \geq 4 \), which holds in practice;
- it is easily shown (by using the definition of a decreasing function and computation for example) that the function \( i \mapsto f(\text{ind}(i), i) \) is decreasing and it’s value in 1 is \( < 1 \) for \( p \geq 3 \). So \( f(\text{ind}(i), i) \leq \text{ind}(i) \) which concludes the induction.

At last we find the final inequality with \( i = q \).

**Proposition A.2.** Algorithm 11 is correct when run with \( \mathcal{P} \)-nonoverlapping expansions.

**Proof:** The proof is similar, the same error analysis holds with the same parameter changes as in Prop. 4.4.

**Bailey’s Iteration Convergence Proof**

**Algorithm 13** Algorithm for reciprocal of the square root of an FP expansion based on Newton-Raphson iteration of the form \( x_{i+1} = x_i + \frac{1}{2} x_i (1 - ax_i^2) \), which is used in QD library

**Input:** FP expansion \( a = a_0 + \ldots + a_{2^q - 1} \); length of output
FP expansion \( 2^q \).

**Output:** FP expansion \( x = x_0 + \ldots + x_{2^q - 1} \) s.t.

$$
\left| x - \frac{1}{\sqrt{a}} \right| \leq \frac{2^{-2^p(p - 3) - 1}}{\sqrt{a}}.
$$

1. \( x_0 = \text{RN}(1/\sqrt{a_0}) \)
2. for \( i \leftarrow 0 \) to \( q - 1 \) do
3. \( \hat{v}[0:2^{i+1} - 1] \leftarrow \text{MulRoundE}(x[0:2^i - 1], a[0:2^{i+1} - 1], 2^{i+1}) \)
4. \( \hat{v}[0:2^{i+1} - 1] \leftarrow \text{MulRoundE}(x[0:2^i - 1], \hat{v}[0:2^{i+1} - 1], 2^{i+1}) \)
5. \( \hat{y}[0:2^{i+1} - 1] \leftarrow \text{SubRoundE}(1, \hat{v}[0:2^{i+1} - 1], 2^{i+1}) \)
6. \( \hat{z}[0:2^{i+1} - 1] \leftarrow \text{MulRoundE}(x[0:2^i - 1], \hat{y}[0:2^{i+1} - 1], 2^{i+1}) \)
7. \( \hat{z}[0:2^{i+1} - 1] \leftarrow \hat{z}[0:2^{i+1} - 1] * 0.5 \)
8. \( x[0:2^{i+1} - 1] \leftarrow \text{AddRoundE}(x[0:2^i - 1], \hat{z}[0:2^{i+1} - 1], 2^{i+1}) \)
9. end for
10. return FP expansion \( x = x_0 + \ldots + x_{2^q - 1} \).

Similarly to the previous proof, one has from Proposition 4.2, eq. (17):

$$
| x_{i+1} - x_i | \leq \gamma_i | x_i | \leq \gamma_i | x_i + \hat{u}_i | \leq \gamma_i | x_i + \frac{\hat{z}_i}{2} ),
$$

$$
| z_i - \hat{z}_i | \leq \gamma_i | z_i | \leq \gamma_i | x_i \hat{y}_i |,
$$

$$
| y_i - \hat{y}_i | \leq \gamma_i | y_i | \leq \gamma_i | 1 - \hat{w}_i |,
$$

$$
| w_i - \hat{w}_i | \leq \gamma_i | w_i | \leq \gamma_i | x_i \hat{v}_i |.
$$
\[|v_i - \hat{v}_i| \leq \gamma_i |v_i| \leq \gamma_i |x_i a(f_i)|, \quad (49)\]

From (45) we have:

\[\varepsilon_{i+1} \leq \left| x_{i+1} - \tau_i \right| + \left| \frac{1}{\sqrt{a}} - \frac{1}{\sqrt{a}} \right| \leq \gamma_i \left| x_i + \frac{\hat{v}_i}{2} \right| + x_i + \frac{\hat{v}_i}{2} - \frac{1}{\sqrt{a}}. \quad (50)\]

Using (46) and (47):

\[\varepsilon_{i+1} \leq \gamma_i (1 + \gamma_i) \left| \frac{1}{2} x_i \hat{y}_i \right| + \gamma_i \left| x_i + \frac{x_i \hat{y}_i}{2} \right| + \left| x_i + \frac{x_i \hat{y}_i}{2} - \frac{1}{\sqrt{a}} \right| \leq \gamma_i (1 + \gamma_i)^2 \left| \frac{1}{2} x_i (1 - \bar{w}_i) \right| + \gamma_i \left| x_i + \frac{1}{2} x_i (1 - \bar{w}_i) + \frac{1}{2} x_i (1 - \bar{w}_i) \right| + \left| x_i + \frac{1}{2} x_i (1 - \bar{w}_i) - \frac{1}{\sqrt{a}} \right| + \gamma_i \left| x_i - \frac{1}{2} x_i (1 - \bar{w}_i) \right| \leq \gamma_i (1 + \gamma_i) (2 + \gamma_i) \left| \frac{1}{2} x_i (1 - \bar{w}_i) \right| + \gamma_i \left| x_i + \frac{1}{2} x_i (1 - \bar{w}_i) \right| + \left| x_i + \frac{1}{2} x_i (1 - \bar{w}_i) - \frac{1}{\sqrt{a}} \right|. \quad (51)\]

By (48) we have:

\[\varepsilon_{i+1} \leq \gamma_i (1 + \gamma_i) (2 + \gamma_i) \left| \frac{1}{2} x_i (1 - \bar{w}_i) \right| + \gamma_i \left| x_i + \frac{1}{2} x_i (1 - \bar{w}_i) \right| + \left| x_i + \frac{1}{2} x_i (1 - \bar{w}_i) - \frac{1}{\sqrt{a}} \right| + \gamma_i (1 + \gamma_i)^3 \left| \frac{1}{2} x_i^2 \hat{v}_i \right|. \]

From (49) we have:

\[\varepsilon_{i+1} \leq \gamma_i (1 + \gamma_i) (2 + \gamma_i) \left| \frac{1}{2} x_i (1 - x_i a(f_i)) \right| + \gamma_i \left| x_i + \frac{1}{2} x_i (1 - x_i^2 a(f_i)) \right| + \left| x_i + \frac{1}{2} x_i (1 - x_i^2 a(f_i)) - \frac{1}{\sqrt{a}} \right| + \gamma_i (1 + \gamma_i)^3 (2 + \gamma_i) \left| \frac{1}{2} x_i^3 a(f_i) \right|. \]

From (50) we have:

\[\varepsilon_{i+1} \leq \gamma_i (1 + \gamma_i) (2 + \gamma_i) \left| \frac{1}{2} x_i (1 - x_i^2 a) \right| + \gamma_i \left| x_i + \frac{1}{2} x_i (1 - x_i^2 a) \right| + \left| x_i + \frac{1}{2} x_i (1 - x_i^2 a) - \frac{1}{\sqrt{a}} \right| + \gamma_i (1 + \gamma_i)^3 (3 + 3 \gamma_i + \gamma_i^2) \left| \frac{1}{2} x_i^3 a \right|. \]

From (42) and similarly to (41), one has:

\[\varepsilon_{i+1} \leq (1 + \gamma_i) \left| x_{i+1} - \frac{1}{\sqrt{a}} \right| + \gamma_i (1 + \gamma_i) (2 + \gamma_i) \left| \frac{1}{2} x_i (1 - x_i^2 a) \right| + \gamma_i (1 + \gamma_i)^3 (3 + 3 \gamma_i + \gamma_i^2) \left| \frac{1}{2} x_i^3 a \right| + \gamma_i \sqrt{a}. \]

We know that \( |x_i \sqrt{a}| \leq \varepsilon_i \sqrt{a} + 1 \) and \( |x_i^3 a| \leq \frac{(\varepsilon_i \sqrt{\varepsilon_i + 1})^3}{\sqrt{a}} \) and moreover \( |\frac{1}{2} x_i (1 - x_i^2 a)| \leq \frac{1}{2} \left( \varepsilon_i + \frac{1}{\sqrt{a}} \right) (a \varepsilon_i + 2 \sqrt{a} \varepsilon_i) \), so we have:

\[\varepsilon_{i+1} \leq \frac{1}{2} (1 + \gamma_i) \sqrt{a} (\varepsilon_i \sqrt{a + 3}) \varepsilon_i^2 + \gamma_i (1 + \gamma_i) (2 + \gamma_i) \left( \frac{\varepsilon_i}{2} \left( \varepsilon_i + \frac{1}{\sqrt{a}} \right) (a \varepsilon_i + 2 \sqrt{a} \varepsilon_i) \right) + \frac{1}{2} \gamma_i (1 + \gamma_i)^3 (3 + 3 \gamma_i + \gamma_i^2) \left( \frac{\varepsilon_i \sqrt{\varepsilon_i + 1})^3}{\sqrt{a}} \right) + \gamma_i \sqrt{a}. \]

Using the notation \( E_i = \varepsilon_i \sqrt{a} \) we can transform (51) in an equation independent of \( a \):

\[E_{i+1} \leq \frac{1}{2} (1 + \gamma_i) (E_i + 3) E_i^2 + \gamma_i (1 + \gamma_i) (2 + \gamma_i) \frac{E_i}{2} (E_i + 1) (E_i + 2) + \frac{1}{2} \gamma_i (1 + \gamma_i)^3 (3 + 3 \gamma_i + \gamma_i^2) (E_i + 1)^3 + \gamma_i. \]

For the last part of the proof, we proceed exactly like in the previous proof and are able to find the same bound \( \text{ind}(i) = 2^{-2(p-3)-1} \) by applying exactly the same inductive reasoning.