Extended-Precision FMA under Parameterized Double-Word Overlap: Tight Error Bounds and Examples

Claude-Pierre Jeannerod, Mioara Joldeş, Nicolas Louvet, Jean-Michel Muller

Extended-Precision FMA under Parameterized Double-Word Overlap: Tight Error Bounds and Examples

Claude-Pierre Jeannerod*, Mioara Joldes**, Nicolas Louvet*, Jean-Michel Muller*
*Inria, UCBL, CNRS, ENSL, LIP, Lyon, France **CNRS, LAAS, Toulouse, France

***Abstract*—We study fast fused multiply–add (FMA) kernels for approximating $ab + c$ in extended precision using double-word (DW) representations in floating-point (FP) arithmetic. We focus on how worst-case guarantees depend on the overlap between the high and low parts, modeled by parameterized conditions of the form $|x_\ell| \leq k\,\mathrm{ulp}(x_h)$ (and two- and three-parameter variants for multiple DW inputs). We consider the dominance regime $|c| \geq 2|ab|$ and its DW analogues (e.g., $|c_h| \geq 2|ab|$ and $|c_h| \geq 2|ab_h|$), which occur in fast paths for elementary-function evaluation and in cancellation-free constructions. Within this regime, we tighten and extend error analyses for several FMA-based schemes used in recent extended-precision implementations. We analyze FP/DW and DW/DW variants, derive explicit worst-case constants for both the returned low part and the overall relative error, and provide matching worst-case examples. The resulting bounds give a unified and quantitative view of how DW-overlap assumptions affect the accuracy of fast extended-precision FMA building blocks.**

***Index Terms*—floating-point arithmetic, double-word arithmetic, fused multiply-add, rounding error analysis; mixed-precision; extended-precision; elementary functions**

I. Introduction

We study fast fused multiply–add (FMA) kernels for approximating $ab + c$ in *extended precision* using *double-word* (DW) representations in floating-point (FP) arithmetic, i.e., pairs of FP numbers whose sum represents a real number more accurately than a simple FP number (see for example [1], [2], [3, Chap. 14] and the references therein). For FP or DW inputs, we seek *worst-case, quantitative guarantees* for the returned DW, and we make explicit how these guarantees depend on admissible overlap between the high and low parts of the input.

The error of a floating-point FMA operation can be expressed exactly as the sum of two FP numbers, and algorithms for doing that are given in [4], but these *general*, exact, algorithms are rather complex.

Another natural baseline is to implement extended-precision updates using *classical* double-word (DW) arithmetic. By composing standard error-free transformations (e.g., TwoSum/FastTwoSum for addition and TwoProdFMA for products), one obtains generic, well-studied kernels for DW addition/multiplication [3, Chap. 14]. These kernels have a non-negligible operation overhead and typically require a final *renormalization* (TwoSum/FastTwoSum-based) step to restore the non-overlapping property.

For implementing $ab + c$, one can use Algorithms DWTimesDW1 and SloppyDWPlusDW which were proposed in the QD Library and analyzed in [2]. Under additional conditions, such as $|c| \gg |ab|$, they can be simplified by skipping a final renormalization step giving very similar algorithms to those introduced in [5] for the so-called pair arithmetic. Furthermore, one can also replace the TwoSum with FastTwoSum. In such case, the classical DW baseline requires on the order of 2–3 multiplications, 1–2 FMAs, and 6–7 additions/subtractions (depending on whether variants like DWTimesDW2 are used).

That said, when $|c| \gg |ab|$, one can design more efficient FMA-centric schemes with controlled worst-case error. Such dominance regimes occur in practice (e.g., in cancellation-free constructions) and arise naturally in polynomial evaluation for elementary functions, as briefly recalled below.

Motivation – FP implementation of elementary functions: Standard implementations first apply a *range reduction* to map the input to a small interval I around 0, using standard functional identities. On I, the function is approximated by a polynomial $P(x) = c_0 + c_1 x + \cdots + c_n x^n$, usually evaluated with Horner's or Estrin's scheme [6]. Let us assume that we use Horner's scheme (although much of the discussion still holds with Estrin's scheme). Early libraries built real-coefficient approximations (e.g., via Remes' algorithm), rounded coefficients to FP, and evaluated the Horner scheme,

$$r_k := c_k + r_{k+1}x, \qquad (n-1 \geq k \geq 0), \qquad r_n := c_n, \quad (1)$$

by successive FMAs, typically achieving an error of a few ulps. When high accuracy is at stake (for example if we aim for correctly rounded functions), we need to use a different strategy. First, correctly-rounded functions [7], [8] are typically implemented using a two-path process: in a *fast path*, a polynomial approximation of reasonably small degree is used, and it is accurate enough that it almost always suffices to determine the correctly rounded result. If the fast path does not suffice, then an *accurate path* is called. In a typical implementation, the accurate path is needed in 0.1% of the cases. Second, we make three remarks.

- Rounding the coefficients of a best (or near-best) real degree-n approximation does *not* in general yield a best (or near-best) degree-n polynomial with FP coefficients. Dedicated methods [9], [10], implemented in Sollya [11], directly target near-best FP polynomials, possibly under additional constraints, e.g., fixing some coefficients, or forcing selected coefficients to be DW.

- Ultimate accuracy is hard to reach when all arithmetic is restricted to the target format. One option is to exploit a wider hardware FP format when available; when it is not, one may resort to DW (or even triple-word) arithmetic, either for intermediate computations and/or to store some coefficients, at a non-negligible cost.
- This cost is nevertheless often acceptable in practice: range reduction makes $|x|$ small and coefficients typically decay rapidly, so high-degree terms contribute little. Thus, fast paths compute the first Horner iterations $r_{n-1}, r_{n-2}, r_{n-3}, \ldots$ *naively* in the target format, and use DW arithmetic only in the last few steps. Accurate paths may use more DW operations, but since they are rarely taken, their contribution to the average latency is small.

This leads to the present focus: the argument is small (typically, $|x| \ll 1$) and polynomial coefficients decrease, so in Horner updates (1) we typically have $|c_k| \gg |r_{k+1}x|$. This important property allows for the design of extended-precision FMA algorithms that are very simple and yet accurate. In particular, for this evaluation strategy, several mixed-precision variants occur: x is either DW (result of the range reduction) or FP (when no reduction is necessary); the partial evaluation r_{k+1} is usually DW (except at the end of the naive part when it is still FP); and, finally, the coefficient c_k is usually FP, becoming DW in the last few critical steps.

Extended-precision FMA cases: We consider the following variants (with their name and corresponding section):

- FP $\times$ FP + FP $\to$ DW [FastTwoFMA, Sec. II-B]
- FP $\times$ FP + DW $\to$ DW [FastTwoFMA_s, Sec. III-A]
- FP $\times$ DW + DW $\to$ DW [FastFMA_DWh, Sec. III-B]
- DW $\times$ DW + DW $\to$ DW [FastFMA_DW, Sec. III-C]

Among these, FastTwoFMA and FastFMA_DW are the most commonly encountered in the above setting. Some algorithms for these schemes were introduced recently in [12]–[15] (FastTwoFMA is called *QuickTwoFMA* in [13]). While [13] targets use cases close to ours (accurate implementations of elementary functions), others consider the application to matrix multiplication [14], [15] or scaled vector addition [12].

Contributions: We give a unified worst-case study for these variants, in the regime where the addend dominates the product. We first tighten the analysis of FastTwoFMA, proving a sharp second-order worst-case bound and giving matching worst-case inputs. We then extend the analysis to the mixed and fully DW kernels, tracking explicitly how allowing overlap between the high and low parts of the input pairs affects: (i) how *small* the returned low part is, and (ii) the overall relative error of the two-term result. We derive explicit worst-case constants (as functions of the overlap parameters) and provide worst-case examples that show these constants are optimal or asymptotically optimal. Finally, we illustrate the practical relevance on a CORE-MATH[1] binary64 `exp` accurate path [7]: expressing the Horner update directly with FastFMA_DW reduces the operation count compared with a classical DW multiply-then-add step, and yields, in our experiments, a measurable speedup of $\simeq 10\%$ on hard-to-round inputs without breaking the accurate-path correctness.

[1]The (anonymous, for blind review) authors are not part of CORE-MATH.

II. Background and preliminaries

We assume a binary, precision-p, FP arithmetic compliant with IEEE-754 (in particular, the arithmetic operations and the FMA are correctly rounded). We denote by $\mathbb{F}$ the set of the FP numbers and, for the sake of simplification, we assume that its exponent range is unbounded. This means that the results presented in this paper hold as soon as underflow and overflow do not occur. In practice, this is not a concern for our problem of evaluating polynomials that approximate functions, as the preliminary range reduction eliminates extreme values. We assume that the rounding function, denoted RN, is round-to-nearest, ties-to-even. The number $u = 2^{-p}$ is the *unit roundoff*, and we frequently use the related quantity $u_1 = u/(1+u)$. If $x \in \mathbb{R}\backslash\{0\}$ then we write $\mathrm{ulp}(x) = 2^{\lfloor \log_2 |x| \rfloor - p + 1}$ and $\mathrm{ufp}(x) = 2^{\lfloor \log_2 |x| \rfloor}$; by convention, $\mathrm{ulp}(0) = \mathrm{ufp}(0) = 0$. Properties of these functions can be found in [3], [16].

A. DW overlap assumptions and dominance regimes

We call double-word (DW) number a pair $x = (x_h, x_\ell)$ such that $x_h = \mathrm{RN}(x_h + x_\ell)$. Whenever convenient, we write $x = x_h + x_\ell$. Such a pair satisfies

$$|x_\ell| \le \tfrac{1}{2}\mathrm{ulp}(x) \le \tfrac{1}{2}\mathrm{ulp}(x_h).$$

(And in particular $x_h = 0$ implies $x_\ell = 0$, since $\mathrm{ulp}(0) = 0$.) We will also consider some relaxed versions, of the form

$$|x_\ell| \le k_x \mathrm{ulp}(x_h), \quad k_x > 0,$$

with k_x being a small FP number or even a small power of two. Similar *overlapping* properties are also considered in [5].

We wish to compute $d := d_h + d_\ell \approx ab + c$ with a, b, c either (possibly relaxed) DW numbers ($a = a_h + a_\ell$, etc.) or FP numbers (e.g., $a_\ell = 0$). The case $|c| \ge 2|ab|$ is of special interest for evaluating elementary functions, which is our target application. We focus on this case in the following, under the name of the *dominance* condition. When we consider "relaxed" DW operands and elementary function evaluation with Horner's rule, a, b, and c do not play a symmetrical role:

- c is a coefficient of a polynomial that has been determined once for all. Hence, it can be renormalized in advance: in all practical cases, we always have $k_c \le \frac{1}{2}$;
- b is the result of the range reduction. It will have the same value in all steps of the Horner evaluation. It therefore makes sense to renormalize it: we very often have $k_b \le \frac{1}{2}$;
- unless there are only a very few Horner steps that need DW arithmetic, it would be costly to perform a renormalization at each step. Hence, in general, $k_a \ge \frac{1}{2}$.

In this context we will heavily use the following properties:

- If $x \in \mathbb{R}$, $f \in \mathbb{F}$, and $|x| \le f$ then $|\mathrm{RN}(x)| \le f$.
- If $x \in \mathbb{R}$ and $r \in \mathbb{R}_{>0}$ then

$$|x| \le r \quad \Rightarrow \quad |\mathrm{RN}(x) - x| \le u \cdot 2^{\lceil \log_2 r \rceil - 1}. \tag{2}$$

- If a is a DW number, then $|a_\ell| \le \min\{u\,\mathrm{ufp}(a), u_1|a|\}$ with $\mathrm{ufp}(a) \le \mathrm{ufp}(a_h) \le |a_h|$.

In addition, the algorithms we examine in this paper all rely on a fourth property, that can be found in [14]: when a, b, c are FP numbers that satisfy the dominance condition ($|c| \geq 2|ab|$) and when $d := \mathrm{RN}(ab + c)$, then $c - d$ is an FP number, i.e., $c - d = \mathrm{RN}(c - d)$. This allows to easily compute, with only one (exact) subtraction and one FMA, the FP number e nearest to the error committed when approximating $ab + c$ by d:

$$e := \mathrm{RN}\big((ab + c) - d\big) = \mathrm{RN}\big(ab + \mathrm{RN}(c - d)\big).$$

B. Improved error analysis of Algorithm FastTwoFMA

We start with a straightforward yet improved analysis of Algorithm 1 FastTwoFMA [14], which given $a, b, c \in \mathbb{F}$ returns $d = (d_h, d_\ell)$ such that the exact sum $d_h + d_\ell$ approximates $ab + c$ with high accuracy under suitable conditions.

Algorithm 1: FastTwoFMA(a, b, c)

$d_h \leftarrow \mathrm{RN}(ab + c)$
$t \leftarrow \mathrm{RN}(c - d_h)$
$d_\ell \leftarrow \mathrm{RN}(ab + t)$
return (d_h, d_ℓ)

As noted in [13], this algorithm is cheap, as it uses only two FMAs and one addition, but can fail to approximate $ab+c$ with about twice the working precision if no additional assumption is made about a, b, c. We refer in particular to [15] for some interesting examples showing that the relative error of $d_h + d_\ell$ can range from about u to 1 and thus be much larger than u^2.

Some conditions were then identified in [14], [15] that suffice to always ensure a relative error in $O(u^2)$. Specifically, it was shown that if the difference $c - d_h$ is in $\mathbb{F}$ then the relative error of $d_h + d_\ell$ is at most $u^2/(1 + u)^2$ and thus less than u^2, and that $|d_\ell| \leq u|ab + c|$; it was also shown that the sufficient condition $c - d_h \in \mathbb{F}$ is satisfied in particular when the dominance condition $|c| \geq 2|ab|$ holds.

In our first result below, we show that the relative error bound $u^2/(1+u)^2$ can in fact be replaced by $\frac{1}{2}u^2$ and that this new bound is tight. We note also that the bound $|d_\ell| \leq u|ab+c|$ from [15] is tight and give a similar one, which is tight as well but involves only the computed quantities d_h and d_ℓ.

Theorem 1. *Given $a, b, c \in \mathbb{F}$, let (d_h, d_ℓ) be the output of* FastTwoFMA *(a, b, c). If $c - d_h \in \mathbb{F}$ then*

$$|d_\ell| \leq \tfrac{1}{2}\mathit{ulp}(d_h) \quad \textit{and} \quad d = (ab + c)(1 + \delta),\ \ |\delta| < \tfrac{1}{2}u^2.$$

Moreover, the bound on $|d_\ell|$ is optimal and the bound on the relative error $|\delta|$ is asymptotically optimal.

Proof. Since $c - d_h \in \mathbb{F}$, we have $d_\ell = \mathrm{RN}(e)$, where $e := ab + c - d_h$. Using $|e| \leq \frac{1}{2}\mathrm{ulp}(ab + c)$ we deduce that $|d_\ell| \leq \frac{1}{2}\mathrm{ulp}(ab + c) \leq \frac{1}{2}\mathrm{ulp}(d_h)$, as wanted.

To bound the relative error we proceed as in the proof of [17, Prop. 1]: since $d-(ab+c) = \mathrm{RN}(e)-e$ with $|e| \leq \frac{1}{2}\mathrm{ulp}(ab+c) = u\mathrm{ufp}(ab+c)$, we have $|d-(ab+c)| \leq \frac{1}{2}u^2\mathrm{ufp}(ab+c)$ as a consequence of (2). If $\mathrm{ufp}(ab+c) < |ab+c|$ then it follows immediately that the relative error is bounded as $|\delta| < \frac{1}{2}u^2$. If $\mathrm{ufp}(ab + c) = |ab + c|$ then $d_h = ab + c$, which implies $t = -ab$, $d_\ell = 0$, and $d_h + d_\ell = ab + c$.

Tightness. If $(a, b, c) = (u, 1, 1)$ then $d_h \in \{1, 1 + 2u\}$, $t = c - d_h \in \{0, -2u\} \subset \mathbb{F}$, $d_\ell = \pm u$, and so $|d_\ell|/\mathrm{ulp}(d_h) = 1/2$ and $|d_\ell|/|ab + c| = u/(1 + u) \sim u$ as $u \to 0$. This means that our bound $|d_\ell| \leq \frac{1}{2}\mathrm{ulp}(d_h)$ is optimal and that the bound $|d_\ell| \leq u|ab + c|$ from [15] is asymptotically optimal.

If $(a, b, c) = (1 - u, \frac{3}{2}u, 1)$ then $ab + c = 1 + \frac{3}{2}u - \frac{3}{2}u^2$, $d_h = 1 + 2u$, $c - d_h = -2u = t$, $d_\ell = -u/2 - 2u^2$, and so $|d - (ab + c)| = \frac{1}{2}u^2$ and $|\delta| \sim \frac{1}{2}u^2$ as $u \to 0$. Hence our relative error bound $\frac{1}{2}u^2$ is asymptotically optimal. □

In practice, the fact that the new bound $\frac{1}{2}u^2$ is tight is easy to observe for the usual FP formats. For example, the input $(a, b, c) = (1-u, \frac{3}{2}u, 1)$ used in the proof above leads to $|\delta| \approx 0.497u^2$ for bf16 arithmetic ($p = 8$) and to $|\delta| \approx 0.499u^2$ for fp16 arithmetic ($p = 11$). This kind of phenomenon can be observed for all the asymptotic bounds we show in this paper.

Note finally that despite the fact that $|d_\ell| \leq \frac{1}{2}\mathrm{ulp}(d_h)$ when $c-d_h \in \mathbb{F}$, the pair (d_h, d_ℓ) returned by FastTwoFMA may *not* be a DW number, that is, it can happen that $d_h \neq \mathrm{RN}(d_h + d_\ell)$. For example, with $p \geq 3$ and $\sigma = p \bmod 2$, taking $a = 1 - \sqrt{u \cdot 2^\sigma/4}$, $b = u(1 + \sqrt{u \cdot 2^\sigma/4})$, and $c = 1 + 2u$ ensures that $a, b, c \in \mathbb{F}$ with $|c| \geq 2|ab|$, and leads to $d_h = 1 + 2u$, $d_\ell = u$, and $\mathrm{RN}(d_h + d_\ell) = \mathrm{RN}(1 + 3u) = 1 + 4u \neq d_h$.

III. DW-overlap analysis of FMA algorithms

We now study how DW overlap impacts worst-case accuracy for the remaining kernels. Each returns $(d_h, d_\ell) \in \mathbb{F}^2$ and we write $d := d_h + d_\ell = (ab+c)(1+\delta)$. The results of our theorems are given in function of the overlap parameters. In particular, Table I summarizes the common case $|x_\ell| \leq \frac{1}{2}\,\mathrm{ulp}(x_h)$ for each DW input. These bounds are (asymptotically) optimal, and matching worst-case inputs are provided.

TABLE I
Worst-case bounds when each DW input x satisfies $|x_\ell| \leq \frac{1}{2}\,\mathrm{ulp}(x_h)$.

Algorithm	$\lvert d_\ell \rvert$	$\lvert\delta\rvert$
FastTwoFMA_s	$\leq \frac{3}{2}\,\mathrm{ulp}(d_h)$	$\leq \dfrac{2u^2}{1-2u}$
FastFMA_DWh	$\leq \frac{5}{2}\,\mathrm{ulp}(d_h)$	$\leq \dfrac{6u^2}{1-4u}$
FastFMA_DW	$\leq 3\,\mathrm{ulp}(d_h)$	$\leq \dfrac{11u^2}{1-6u-u^2}$

A. FastTwoFMA_s *(FP × FP + DW → DW)*

We consider here Algorithm 2 FastTwoFMA_s as in [14], with a and b two FP numbers and $c = c_h + c_\ell$ a DW number.

Assuming $|c_h| \geq 2|ab|$, two rigorous error bounds are established in [15, Theorem 6], which read essentially as

$$|d_\ell| \lesssim u|ab + c_h| + |c_\ell|,$$

and, writing $d = d_h + d_\ell$,

$$|d - (ab + c)| \lesssim 2u^2|ab + c_h| + u|c_\ell|.$$

Algorithm 2: FASTTWOFMA_S(a, b, c_h, c_ℓ)

$d_h \leftarrow \mathrm{RN}(ab + c_h)$
$t \leftarrow \mathrm{RN}(c_h - d_h)$
$e \leftarrow \mathrm{RN}(ab + t)$
$d_\ell \leftarrow \mathrm{RN}(e + c_\ell)$
return $(d_h,\, d_\ell)$

We note first that the bound on $|d_\ell|$ is asymptotically optimal. For example, if $(a, b, c_h, c_\ell) = (-\frac{1}{2}+\frac{1}{2}u, 1, 1, u-u^2)$ then $a, b \in \mathbb{F}$, c is a DW, and $|c_h| \geq 2|ab|$. Furthermore, for $p \geq 3$, $d_\ell = \frac{3}{2}u$ and $u|ab + c_h| + |c_\ell| = \frac{3}{2}u - \frac{1}{2}u^2$, so that both quantities are asymptotically equivalent as $u \to 0$.

The second bound, however, can be refined easily by applying Theorem 1, as the next lemma shows.

Lemma 1. *Given $a, b \in \mathbb{F}$ and $c = (c_h, c_\ell)$ a DW number, let $d = (d_h, d_\ell)$ be the output of* FASTTWOFMA_S(a, b, c_h, c_ℓ). *If $|c_h| \geq 2|ab|$ then*

$$|d - (ab + c)| \leq \tfrac{3}{2}u^2|ab + c_h| + u|c_\ell|.$$

Proof. We have $|d_h + d_\ell - (ab + c_h + c_\ell)| \leq A + A'$ with $A = |d_h + e - (ab + c_h)| \leq \frac{1}{2}u^2|ab + c_h|$ by Theorem 1, and $A' = |d_\ell - (e + c_\ell)| \leq u_1|e + c_\ell| \leq u|e| + u|c_\ell|$ with $|e| = |ab + c_h - d_h| \leq u_1|ab + c_h| \leq u|ab + c_h|$. □

Using this bound on the absolute error together with the condition $|ab| \leq |c_h|/2$ and the fact that $|c_\ell| \leq u|c_h|$ yields

$$|d_h + d_\ell - (ab + c)| \leq \tfrac{13}{4}u^2|c_h|,$$
$$|ab + c| \geq |c_h| - |ab| - |c_\ell| \geq \left(\tfrac{1}{2} - u\right)|c_h|,$$

so that $d = (ab + c)(1 + \delta)$ with $|\delta| \leq \frac{13}{2-4u}u^2 \sim 6.5u^2$.

However, this bound on the relative error is not tight. This is a direct consequence of our Theorem 2 below, which proves an optimal bound on $|d_\ell|/\mathrm{ulp}(d_h)$ and an asymptotically optimal bound on $|\delta|$, each as a function of the overlap parameter k_c. In particular, when $k_c = 1/2$ (which is the case when c is a DW number), this theorem tells us that a tight bound on $|\delta|$ is $\frac{2u^2}{1-2u}$, which is $\sim 2u^2$ as $u \to 0$, and thus more than three times smaller than the previous bound $\sim 6.5u^2$.

Theorem 2. *Let $a, b \in \mathbb{F}$ and $c = (c_h, c_\ell) \in \mathbb{F}^2$ be such that*

$$|c_\ell| \leq k_c\, \mathit{ulp}(c_h), \qquad k_c \in \mathbb{R}_{>0},$$

and let $(d_h, d_\ell) :=$ FASTTWOFMA_S(a, b, c_h, c_ℓ).

If $|c_h| \geq 2|ab|$ then

$$|d_\ell| \leq \frac{4k_c + 1}{2}\, \mathit{ulp}(d_h)$$

and $d = (ab + c)(1 + \delta)$ with

$$|\delta| \leq \frac{\alpha_c\, u^2}{1 - 4k_c u}, \qquad \alpha_c := 2^{\lceil \log_2(4k_c+1)\rceil - 1},$$

where we assume[2] *that $4k_c + 1 \in \mathbb{F}$. Furthermore, the bound on $|d_\ell|$ is optimal and the one on $|\delta|$ is asymptotically optimal.*

[2]This assumption, which is for mostly simplicity, is reasonable in practice: as said in §II, k_c is often a small FP number or even a small power of two.

Proof. For simplicity, let us write $k = k_c$ and $x = ab + c_h$. Since $c_h - d_h \in \mathbb{F}$ when $|c_h| \geq 2|ab|$, we have $d_h = \mathrm{RN}(x)$ and $e = \mathrm{RN}(x - d_h)$.

If $x = 0$ then $|c_h| = |ab| \geq 2|ab|$, so $ab = c_h = 0$. Hence $c_\ell = 0$ and at least one of a and b is zero, which implies that the exact result $ab + c_h + c_\ell$ is zero. Now, $d_h = d_\ell = 0$, and so the bounds claimed for $|d_\ell|$ and $|\delta|$ hold in this case.

Assume that $x \neq 0$ and, up to scaling and changing the signs of b and c, that $x \in [1, 2)$. Defining $\epsilon_1 := e - (x - d_h)$ and $\epsilon_2 := d_\ell - (e + c_\ell)$, we have $d - (ab + c) = \epsilon_1 + \epsilon_2$.

Since $x \in [1, 2)$, we have $d_h \in [1, 2]$, $\mathrm{ulp}(d_h) \in \{2u, 4u\}$, $|x - d_h| \leq u$, $|e| \leq u$, and $|\epsilon_1| \leq \frac{1}{2}u^2$.

Using $|c_h| \geq 2|ab|$, we deduce that $2 > ab + c_h \geq |c_h| - |ab| \geq |c_h|/2$ and thus $|c_h| < 4$ and $|ab| < 2$. It follows that $\mathrm{ulp}(c_h) \leq 4u$, which gives $|c_\ell| \leq 4ku$ and thus $|e + c_\ell| \leq (4k + 1)u$. Since $4k + 1 \in \mathbb{F}$ and $\mathrm{ulp}(d_h) \geq 2u$, we obtain $|d_\ell| \leq (4k + 1)u \leq (4k + 1)/2 \cdot \mathrm{ulp}(d_h)$.

Let us now bound the relative error $|\delta| = |d/(ab+c)-1|$. If $|c_h| < 2$ then $\mathrm{ulp}(c_h) \leq 2u$, $|c_\ell| \leq 2ku$, $|e + c_\ell| \leq (2k + 1)u$, and, using (2), we deduce that

$$|\epsilon_2| \leq \lambda_k u^2, \qquad \lambda_k := 2^{\lceil \log_2(2k+1)\rceil - 1}.$$

Hence $|d - (ab + c)| \leq |\epsilon_1| + |\epsilon_2| \leq (\frac{1}{2} + \lambda_k)u^2$ and, using the fact that $|ab + c| \geq |x| - |c_\ell| \geq 1 - 2ku$, we deduce that

$$|\delta| \leq \frac{\frac{1}{2} + \lambda_k}{1 - 2ku}u^2 =: B_1 \qquad \text{if } |c_h| < 2.$$

Assume now that $|c_h| \in [2, 4)$. Then $\mathrm{ulp}(c_h) \leq 4u$, $|c_\ell| \leq 4ku$, $|e + c_\ell| \leq (4k + 1)u$, and $|\epsilon_2| \leq \lambda_{2k}u^2$.

- If $|ab| < 1/2$ then $|ab+c| \geq |c_h|-|ab|-|c_\ell| \geq 3/2-4ku$. Therefore, recalling that $|\epsilon_1| \leq \frac{1}{2}u^2$, we obtain
$$|\delta| \leq \frac{\frac{1}{2} + \lambda_{2k}}{\frac{3}{2} - 4ku}u^2 =: B_2 \qquad \text{if } |c_h| \geq 2 \text{ and } |ab| < 1/2.$$
- If $|ab| \geq 1/2$ then, since $a, b \in \mathbb{F}$, one can check that $ab \in u^2\mathbb{Z}$ and, using $|c_h| \geq 2$, that $x - d_h \in u^2\mathbb{Z}$ as well. Since $|x - d_h| \leq u$, we deduce that $x - d_h \in \mathbb{F}$ and thus $\epsilon_1 = 0$. Hence $|d - (ab + c)| = |\epsilon_2|$ and it follows from $|ab + c| \geq |x| - |c_\ell| \geq 1 - 4ku$ that
$$|\delta| \leq \frac{\lambda_{2k}}{1 - 4ku}u^2 =: B_3 \qquad \text{if } |c_h| \geq 2 \text{ and } |ab| \geq 1/2.$$

To conclude, it suffices to check that $B_1 \leq B_3$ (by using the fact that $\lambda_k \leq 2k \leq \lambda_{2k} - 1/2$) and $B_2 \leq B_3$ (by using the fact that $k > 0$ implies $\lambda_{2k} \geq 1 \geq 1 - 4ku$).

Tightness. Taking $(a, b, c_h, c_\ell) = (-1, 1 - u, 2, 4ku)$ gives $\mathrm{ulp}(d_h) = 2u$ and $d_\ell = (4k+1)u$, which shows that our bound on $|d_\ell|$ can be attained. However, $d = ab + c \neq 0$ in this case, and so the relative error δ is zero. To show that our bound on $|\delta|$ is tight, we take instead $(a, b, c_h, c_\ell) = (-1, 1-u, 2, u(1-u)\alpha_c)$. In this case, $|c_\ell| \leq u(1 - u) \cdot 4k < 4ku = k\mathrm{ulp}(c_h)$, $d_h = 1$, $e = u$, $\epsilon_1 = 0$, and $e + c_\ell = (1 + 2^{-\ell} - u) \cdot 2^\ell u$ with $\ell := \lceil \log_2(4k + 1)\rceil - 1$. For $k > 0$ and $4k + 1 \leq 1/u$, we have $0 \leq \ell \leq p - 1 = -\log_2 u - 1$ and thus $1 + 2^{-\ell} \in \mathbb{F} \cap (1, 2]$. Hence $|\epsilon_2| = 2^\ell u^2 = \alpha_c u^2$ and so $|d/(ab+c)-1| = \alpha_c u^2/(1 + u + u(1 - u)\alpha_c)$, which is equivalent to the upper bound $\alpha_c u^2/(1 - 4ku)$ for fixed k and as $u \to 0$. □

B. FastFMA_DWH *(FP × DW + DW → DW)*

Algorithm 3 FastFMA_DWH was given in [15], without an error analysis. Of course, our analysis of Algorithm 4 with $a_\ell = 0$ applies, but in the following we establish better bounds.

Algorithm 3: FastFMA_DWH$(a,\, b_h,\, b_\ell,\, c_h,\, c_\ell)$

$d_h \leftarrow \mathrm{RN}(ab_h + c_h)$
$t \leftarrow \mathrm{RN}(c_h - d_h)$
$e \leftarrow \mathrm{RN}(ab_h + t)$
$f \leftarrow \mathrm{RN}(e + c_\ell)$
$d_\ell \leftarrow \mathrm{RN}(ab_\ell + f)$
return $(d_h,\, d_\ell)$

Our bounds are parameterized by the overlap factors k_b and k_c of b and c. In particular, when b and c are DW numbers, we can take $k_b = k_c = 1/2$, and in this case Theorem 3 gives the tight bounds $|d_\ell|/\mathrm{ulp}(d_h) \le 5/2$ and $|\delta| \le 6u^2/(1-4u)$.

Theorem 3. *Let $a \in \mathbb{F}$ and $b = (b_h, b_\ell), c = (c_h, c_\ell) \in \mathbb{F}^2$ be such that*

$$|b_\ell| \le k_b\, ulp(b_h) \quad \text{and} \quad |c_\ell| \le k_c\, ulp(c_h)$$

for some constants $k_b, k_c \in \mathbb{R}_{>0}$, and let (d_h, d_ℓ) be the output of FastFMA_DWH$(a,\, b_h,\, b_\ell,\, c_h,\, c_\ell)$.

If $|c_h| \ge 2|ab_h|$ then

$$|d_\ell| \le \frac{4k_b + 4k_c + 1}{2}\, ulp(d_h)$$

and $d = (ab + c)(1 + \delta)$ with

$$|\delta| \le \max\left\{ \frac{(\frac{1}{2} + \alpha_c + \alpha_{b,c})u^2}{1 - (2k_b + 4k_c)u}, \frac{(\alpha_c + \alpha'_{b,c})u^2}{1 - (4k_b + 4k_c)u} \right\},$$

where α_c is as in Theorem 2, $\alpha_{b,c}$ and $\alpha'_{b,c}$ are defined as

$$\alpha_{b,c} := 2^{\lceil \log_2(2k_b + 4k_c + 1) \rceil - 1},$$
$$\alpha'_{b,c} := 2^{\lceil \log_2(4k_b + 4k_c + 1) \rceil - 1},$$

and the constants k_b, k_c, $4k_c{+}1$, and $4k_b{+}4k_c{+}1$ are assumed to be in $\mathbb{F}$. Moreover, the bound on $|d_\ell|$ is asymptotically optimal and, when $k_b = k_c = 1/2$, the bound on $|\delta|$ is asymptotically optimal as well.

Proof. Let $x = ab_h + c_h$. If $x = 0$ then one proceeds as for FastTwoFMA_S. If $x \neq 0$ then, up to scaling and changing the signs of b_h and c_h, one assumes $x \in [1, 2)$. This implies $2 > |ab_h + c_h| \ge |c_h| - |ab_h|$, which for $|ab_h| \le |c_h|/2$ gives $|c_h| < 4$ and $|ab_h| < 2$.

From $x \in [1, 2)$, $|ab_h| < 2$, and $|c_h| < 4$ it follows that $d_h \in [1, 2]$, $|x - d_h| \le u$, $|e| \le u$, $|c_\ell| \le k_c\, \mathrm{ulp}(c_h) \le 4k_c u$, and $|ab_\ell| \le |a| \cdot k_b\, \mathrm{ulp}(b_h) \le 2k_b u |ab_h| \le 4k_b u$. Hence $|e + c_\ell| \le (4k_c + 1)u$ and since this bound is in $\mathbb{F}$, we have $|f| \le (1{+}4k_c)u$ as well. Then $|ab_\ell + f| \le (4(k_b{+}k_c){+}1)u \in \mathbb{F}$ and thus $|d_\ell| \le (4(k_b + k_c) + 1)u$ as well. Since $d_h \in [1, 2]$ implies $\mathrm{ulp}(d_h) \ge 2u$, we deduce that $|d_\ell| \le (4k_b + 4k_c + 1)/2 \cdot \mathrm{ulp}(d_h)$.

To bound the relative error $|\delta| = |d/(ab + c) - 1|$, define $\epsilon_1 := e - (x - d_h)$, $\epsilon_2 := f - (e + c_\ell)$, and $\epsilon_3 := d_\ell - (ab_\ell + f)$. Since $|x - d_h| \le u$ and $|e + c_\ell| \le (4k_c + 1)u$, using (2) gives $|\epsilon_1| \le \frac{1}{2}u^2$ and $|\epsilon_2| \le \alpha_c u^2$. Let us now bound $|\epsilon_3|$.

If $|ab_h| \le 1$ then $|ab_\ell| \le |a| \cdot k_b \mathrm{ulp}(b_h) \le 2k_b u$, so that $|ab_\ell + f| \le (2k_b + 4k_c + 1)u$ and thus $|\epsilon_3| \le \alpha_{b,c} u^2$. Since $|ab + c| \ge |x| - |ab_\ell| - |c_\ell| \ge 1 - 2k_b u - 4k_c u$, we obtain

$$|\delta| \le \frac{\frac{1}{2} + \alpha_c + \alpha_{b,c}}{1 - (2k_b + 4k_c)u} u^2 =: B \qquad \text{if } |ab_h| \le 1.$$

If $|ab_h| > 1$ then $ab_h \in 2u^2\mathbb{Z}$ and so is x, which together with $|x - d_h| \le u$ implies $x - d_h \in \mathbb{F}$, that is, $\epsilon_1 = 0$. Hence $|d - (ab + c)| \le |\epsilon_2| + |\epsilon_3| \le \alpha_c u^2 + |\epsilon_3|$. Since $|ab_h| < 2$, we have $|ab_\ell| \le |a| \cdot k_b \mathrm{ulp}(b_h) \le 4k_b u$, so that $|ab_\ell + f| \le (4k_b + 4k_c + 1)u$ and thus $|\epsilon_3| \le \alpha'_{b,c} u^2$. Since $|ab + c| \ge |x| - |ab_\ell| - |c_\ell| \ge 1 - 4k_b u - 4k_c u$, we conclude that

$$|\delta| \le \frac{\alpha_c + \alpha'_{b,c}}{1 - (4k_b + 4k_c)u} u^2 =: B' \qquad \text{if } |ab_h| > 1.$$

We thus have shown that $|\delta| \le \max\{B, B'\}$.

Note that B can be larger or smaller than B' depending on the choice of parameters (k_b, k_c). For example, $k_b = k_c = 1/2$ gives $B = 4.5u^2/(1 - 3u) < B' = 6u^2/(1 - 4u)$, while $(k_b, k_c) = (1/2, 1)$ gives $B = 8.5u^2/(1 - 5u)$ and $B' = 8u^2/(1 - 6u)$.

Tightness of the bound on $|d_\ell|$. For p even such that $p \ge 6$, let $a = -2 + 3\sqrt{u} + 12u$, $(b_h, b_\ell) = (1 + \sqrt{u} - 4u, -2k_b u)$, and $(c_h, c_\ell) = (4 - \sqrt{u} - 24u, 4k_c u)$. One can check that $a, b_h, b_\ell, c_h, c_\ell \in \mathbb{F}$, $|b_\ell| = k_b \mathrm{ulp}(b_h)$, $|c_\ell| = k_c \mathrm{ulp}(c_h)$, and $|ab_h| \le |c_h|/2$. Furthermore, $x = 2 - u - 48u^2$ is rounded down to $d_h = 2 - 2u$, so $\mathrm{ulp}(d_h) = 2u$ and $e = u - 48u^2$. Hence $e + c_\ell = u(4k_c + 1 - 48u)$ and, therefore, $f = (4k_c + 1)u + O(u^2)$. Since $ab_\ell = (2 - 3\sqrt{u} - 12u) \cdot 2k_b u = 4k_b u + O(u^{3/2})$, we deduce that $d_\ell = (4k_b + 4k_c + 1)u + O(u^{3/2})$ and, therefore, $|d_\ell|/\mathrm{ulp}(d_h) \sim (4k_b + 4k_c + 1)/2$ as $u \to 0$.

For p odd such that $p \ge 7$, a similar conclusion can be reached by considering $a = -2 + \sqrt{2u} + 8u$, $b_h = 1 + \frac{1}{2}\sqrt{2u} - 4u$, $c_h = 4 - 24u$, and b_ℓ, c_ℓ as before.

Tightness of the bound on $|\delta|$ when $k_b = k_c = 1/2$. In this case the relative error bound is $|\delta| \le 6u^2/(1 - 4u)$. To show that it is asymptotically optimal, consider $a = -(1 + \sqrt{u \cdot 2^\sigma})$ with $\sigma = p \bmod 2$, $(b_h, b_\ell) = (1 + \sqrt{u/2^\sigma}, u - 4u^2)$, and $(c_h, c_\ell) = (2 + 4\sqrt{u \cdot 2^\sigma} + 4u, -2u + 6u^2)$. Then one can check that $a, b \in \mathbb{F}$, c is a DW number, and $|c_h| \ge 2|ab_h|$. Furthermore, for $p \ge 8$ even, $d_h = 1 + 2\sqrt{u} + 4u$ and $d_\ell = -4u - u^{3/2} + 1 - u^2$, so that $d = 1 + 2\sqrt{u} - u^{3/2} + 16u^2$, which together with $ab + c = 1 + 2\sqrt{u} - u^{3/2} + 10u^2 + 4u^{5/2}$ shows that the relative error $|d/(ab + c) - 1|$ is $6u^2 - O(u^{5/2}) \sim 6u^2$.

The case where p is odd can be checked in the same way: for $p \ge 7$, $d_h = 1 + \frac{5}{2}\sqrt{2u} + 4u$, $e = -u$, $f = -3u + 8u^2$, and $d_\ell = -4u - \sqrt{2u} \cdot u + 16u^2$, so that $d = 1 + (\frac{5}{2} - u)\sqrt{2u} + 16u^2$, $ab + c = 1 + (\frac{5}{2} - u)\sqrt{2u} + 10u^2 + 4u^2\sqrt{2u}$, and thus $|d/(ab + c) - 1| = 6u^2 - 19\sqrt{2}u^{5/2} + O(u^3) \sim 6u^2$. □

C. FastFMA_DW *(DW × DW + DW → DW)*

Here a, b, c are DW numbers and an approximation $d \approx ab + c$ is computed following [15] according to Algorithm 4 FastFMA_DW.

Algorithm 4: FASTFMA_DW$(a_h,\, a_\ell,\, b_h,\, b_\ell,\, c_h,\, c_\ell)$

$d_h \leftarrow \mathrm{RN}(a_h b_h + c_h)$
$t \leftarrow \mathrm{RN}(c_h - d_h)$
$e \leftarrow \mathrm{RN}(a_h b_h + t)$
$f \leftarrow \mathrm{RN}(e + c_\ell)$
$g \leftarrow \mathrm{RN}(a_h b_\ell + f)$
$d_\ell \leftarrow \mathrm{RN}(a_\ell b_h + g)$
return $(d_h,\, d_\ell)$

Assuming $|c_h| \geq 2|a_h b_h|$, two rigorous error bounds are given for this algorithm in [15]. Up to higher-order terms in u, the first one has the form $|d_\ell| \lesssim 3u|a_h b_h| + u|c_h| + |c_\ell|$ and can in fact easily be replaced by

$$|d_\ell| \lesssim 2u|a_h b_h| + u|a_h b_h + c_h| + |c_\ell|.$$

This bound is tight: taking $(a_h, a_\ell) = (-1-2u, -u+u^2)$, $(b_h, b_\ell) = (1, u)$, and $(c_h, c_\ell) = (3+8u, -u)$ yields a ratio $|d_\ell|/\text{bound} = 5u/(5u+10u^2)$ that tends to 1 as $u \to 0$.

The second bound from [15] is

$$|d - (ab+c)| \leq 8u^2|a_h b_h| + 4u^2|c_h| + 3u|c_\ell|,$$

and it turns out that it can be refined into

$$|d - (ab+c)| \leq \tfrac{15}{2}u^2|a_h b_h| + \tfrac{7}{2}u^2|c_h| + 3u|c_\ell|$$

by a direct application of our Lemma 1 to the analysis done in [15, p. 12]. Combining this new bound with the inequalities $|a_h b_h| \leq |c_h|/2$ and $|y_\ell| \leq u|y_h|$ for $y \in \{a, b, c\}$ would give

$$\begin{aligned} |d - (ab+c)| &\leq 10.25u^2|c_h|, \\ |ab+c| &\geq |c_h| - |a_h b_h| - |a_h b_\ell| - |a_\ell b_h| - |a_\ell b_\ell| - |c_\ell| \\ &\geq \left(\tfrac{1}{2} - 2u - \tfrac{1}{2}u^2\right)|c_h|, \end{aligned}$$

and thus a relative error bound $|\delta| \lesssim 20.5u^2$.

This bound can be replaced by the twice smaller and provably tight bound $|\delta| \leq 11u^2/(1-6u-u^2)$, which Theorem 4 below yields directly when setting $k_a = k_b = k_c = 1/2$ for the overlap parameters of a, b, and c. This is just one example of application of this theorem, which allows for considerable flexibility in the choice of these three parameters.

Another example will be given in Corollary 1 for the case $k_a \geq k_b = k_c = 1/2$, which, as seen in Sec. I, is particularly relevant in the context of DW Horner evaluation and for which the asymptotically optimal bound $|d_\ell| \leq (2k_a+2)\mathrm{ulp}(d_h)$ can then be derived easily.

Theorem 4. *Let $a = (a_h, a_\ell), b = (b_h, b_\ell), c = (c_h, c_\ell) \in \mathbb{F}^2$ be such that*

$$|a_\ell| \leq k_a \mathit{ulp}(a_h), \quad |b_\ell| \leq k_b \mathit{ulp}(b_h), \quad |c_\ell| \leq k_c \mathit{ulp}(c_h)$$

for some constants $k_a, k_b, k_c \in \mathbb{R}_{>0}$, and let (d_h, d_ℓ) be the output of FastFMA_DW$(a_h, a_\ell, b_h, b_\ell, c_h,\, c_\ell)$.

If $|c_h| \geq 2|a_h b_h|$ then

$$|d_\ell| \leq \begin{cases} 3\,\mathit{ulp}(d_h) & \text{if } k_a = k_b = k_c = 1/2, \\ \frac{4k_a+4k_b+4k_c+1}{2}\,\mathit{ulp}(d_h) & \text{otherwise,} \end{cases}$$

and the relative error is bounded as

$$|\delta| \leq \max\{B, B'\},$$

where

$$B := \frac{\frac{1}{2} + \alpha_c + \alpha_{b,c} + \alpha_{a,b,c} + 4k_a k_b}{1 - (2k_a + 2k_b + 4k_c)u - 4k_a k_b u^2} u^2$$

and

$$B' := \frac{\alpha_c + \alpha'_{b,c} + \alpha'_{a,b,c} + 4k_a k_b}{1 - (4k_a + 4k_b + 4k_c)u - 4k_a k_b u^2} u^2.$$

Here, α_c, $\alpha_{b,c}$, and $\alpha'_{b,c}$ are as in Theorems 2 and 3, $\alpha_{a,b,c}$ and $\alpha'_{a,b,c}$ are defined as

$$\begin{aligned} \alpha_{a,b,c} &:= 2^{\lceil \log_2(2k_a+2k_b+4k_c+1)\rceil - 1}, \\ \alpha'_{a,b,c} &:= 2^{\lceil \log_2(4k_a+4k_b+4k_c+1)\rceil - 1}, \end{aligned}$$

and $4k_c+1$, $2k_b+4k_c+1$, $4k_b+4k_c+1$, $2k_a+2k_b+4k_c+1$, and $4k_a+4k_b+4k_c+1$ are assumed to be in $\mathbb{F}$.

Moreover, when $k_a = k_b = k_c = 1/2$, these bounds on $|d_\ell|$ and $|\delta|$ are asymptotically optimal.

Proof. Let $x = a_h b_h + c_h$. As for the two previous theorems, we can handle the case $x = 0$ separately, and then assume $x \in [1, 2)$, which gives $|c_h| < 4$, $|a_h b_h| < 2$, and $|c_\ell| \leq 4k_c u$.

It follows that $\mathrm{ufp}(a_h)\mathrm{ufp}(b_h) \leq 1$, which together with $|a_\ell b_\ell| \leq k_a \mathrm{ulp}(a_h) \cdot k_b \mathrm{ulp}(b_h)$ gives $|a_\ell b_\ell| \leq 4k_a k_b u^2$. Furthermore, the two other products $a_h b_\ell$ and $a_\ell b_h$ satisfy

$$|a_h b_\ell| \leq 2k_b u|a_h b_h| \quad \text{and} \quad |a_\ell b_h| \leq 2k_a u|a_h b_h|.$$

Let ϵ_1, ϵ_2, ϵ_3, and ϵ_4 be defined by $\epsilon_1 = e - (x - d_h)$, $\epsilon_2 = f - (e + c_\ell)$, $\epsilon_3 = g - (a_h b_\ell + f)$, and $\epsilon_4 = d_\ell - (a_\ell b_h + g)$. The absolute error is then bounded as

$$|d - (ab+c)| \leq |\epsilon_1| + |\epsilon_2| + |\epsilon_3| + |\epsilon_4| + |a_\ell b_\ell|.$$

A bound on the relative error $|\delta| = |d/(ab+c) - 1|$ will then follow immediately by using further the lower bound

$$\begin{aligned} |ab+c| &\geq |x| - |a_h b_\ell| - |a_\ell b_h| - |a_\ell b_\ell| - |c_\ell| \\ &\geq 1 - (2k_a u + 2k_b u)|a_h b_h| - 4k_c u - 4k_a k_b u^2. \end{aligned}$$

Since $x \in [1, 2)$, we have $d_h \in [1, 2)$, $\mathrm{ulp}(d_h) \geq 2u$, $|e| \leq u$, and $|\epsilon_1| \leq \frac{1}{2}u^2$. Using $|c_\ell| \leq 4k_c u$ then yields as before $|e + c_\ell| \leq (4k_c+1)u$ and $|\epsilon_2| \leq \alpha_c u^2$, and since $4k_c + 1 \in \mathbb{F}$, we obtain $|f| \leq (4k_c+1)u$. Let us now bound $|\epsilon_3|$ and $|\epsilon_4|$.

If $|a_h b_h| \leq 1$ then $|a_h b_\ell + f| \leq |a_h| \cdot 2k_b u|b_h| + |f| \leq (2k_b + 4k_c + 1)u$, $|\epsilon_3| \leq \alpha_{b,c} u^2$, and, using the fact that $2k_b + 4k_c + 1$ is in $\mathbb{F}$, we deduce that $|g| \leq (2k_b + 4k_c + 1)u$. Then $|a_\ell b_h + g| \leq 2k_a u|a_h| \cdot |b_h| + |g| \leq (2k_a + 2k_b + 4k_c + 1)u$, $|\epsilon_4| \leq \alpha_{a,b,c} u^2$, and, since $2k_a + 2k_b + 4k_c + 1$ is $\mathbb{F}$, we obtain

$$\begin{aligned} |d_\ell| &\leq (2k_a + 2k_b + 4k_c + 1)u \\ &\leq (k_a + k_b + 2k_c + 1/2)\mathrm{ulp}(d_h). \end{aligned}$$

It also follows that the relative error $|\delta|$ is bounded as

$$|\delta| \leq \frac{(\frac{1}{2} + \alpha_c + \alpha_{b,c} + \alpha_{a,b,c} + 4k_a k_b)u^2}{1 - (2k_a + 2k_b + 4k_c)u - 4k_a k_b u^2} =: B.$$

If $|a_h b_h| > 1$ then $a_h b_h \in 2u^2\mathbb{Z}$, so that $u \geq |a_h b_h + c_h - d_h| \in 2u^2\mathbb{Z}$, which implies $\epsilon_1 = 0$ and thus $|d - (ab + c)| \leq \alpha_c u^2 + |\epsilon_3| + |\epsilon_4| + 4k_a k_b u^2$. Now, $|a_h b_\ell + f| \leq 2k_b u|a_h b_h| + |f| \leq (4k_b + 4k_c + 1)u$ and thus $|\epsilon_3| \leq \alpha'_{b,c} u^2$. Since $4k_b + 4k_c + 1$ is in $\mathbb{F}$, we deduce also that $|g| \leq (4k_b + 4k_c + 1)u$ and thus, recalling that $|a_h b_h| < 2$, $|a_\ell b_h + g| \leq 2k_a u|a_h b_h| + |g| \leq (4k_a + 4k_b + 4k_c + 1)u$. Hence $|\epsilon_4| \leq \alpha'_{a,b,c} u^2$ and, using the fact that $4k_a + 4k_b + 4k_c + 1$ is in $\mathbb{F}$,

$$\begin{aligned} |d_\ell| &\leq (4k_a + 4k_b + 4k_c + 1)u \\ &\leq (2k_a + 2k_b + 2k_c + 1/2)\mathrm{ulp}(d_h). \end{aligned}$$

Furthermore, the relative error $|\delta|$ is now bounded as

$$|\delta| \leq \frac{(\alpha_c + \alpha'_{b,c} + \alpha'_{a,b,c} + 4k_a k_b)u^2}{1 - (4k_a + 4k_b + 4k_c)u - 4k_a k_b u^2} =: B'.$$

From these two subcases, we conclude that the relative error is bounded as $|\delta| \leq \max\{B, B'\}$. Here, B can be larger or smaller than B' depending on the values of the parameters k_a, k_b, k_c. For example, $B \sim 9.5u^2$ and $B' \sim 11u^2$ when $k_a = k_b = k_c = 1/2$, while $B \sim 25.5u^2$ and $B' \sim 25u^2$ when $k_a = k_b = 1/2$ and $k_c = 2$. For the bound on $|d_\ell|$, the second bound shown above is always larger than the first one, which concludes the analysis for the general case.

Refinement when $k_a = k_b = k_c = 1/2$. For such values of the parameters the bounds obtained so far on $|d_\ell|$ are $\frac{5}{2}\mathrm{ulp}(d_h)$ if $|a_h b_h| \leq 1$, and $\frac{7}{2}\mathrm{ulp}(d_h)$ if $|a_h b_h| > 1$. Let us now show that in this second case, the smaller bound $3\mathrm{ulp}(d_h)$ holds. First, we have $|f| \leq 3u$ and thus $|g| \leq (|a_h b_\ell| + 3u)(1 + u_1)$. It follows that

$$\begin{aligned} |a_\ell b_h + g| &\leq |a_\ell b_h| + |g| \\ &\leq |a_\ell b_h| + |a_h b_\ell|(1 + u_1) + 3u(1 + u_1). \end{aligned}$$

Up to scaling (a_h, a_ℓ) and (b_h, b_ℓ), we assume $|a_h| \in [1, 2)$. This implies $|a_\ell| \leq u$, $|b_h| < 2/|a_h| \leq 2$, and $|b_\ell| \leq u$. Hence

$$|a_\ell b_h + g| \leq \frac{2u}{|a_h|} + u(1 + u_1)|a_h| + 3u(1 + u_1) =: \varphi(|a_h|).$$

For $|a_h| \in [1, 2-2u]$, it can be checked that $\varphi(|a_h|) \leq \varphi(1) = 2u + 4u(1 + u_1)$, which is strictly less than the midpoint $6u + 4u^2$. Hence, $|d_\ell| = |\mathrm{RN}(a_\ell b_h + g)| \leq 6u$. Recalling that $\mathrm{ulp}(d_h) \geq 2u$, we arrive at $|d_\ell| \leq 3\mathrm{ulp}(d_h)$, as wanted.

Asymptotic optimality when $k_a = k_b = k_c = 1/2$. We have in this case $|d_\ell| \leq 3\mathrm{ulp}(d_h)$ and $|\delta| \leq \max\{B, B'\} = B' = 11u^2/(1 - 6u - u^2)$.

To show that this bound on $|d_\ell|$ is asymptotically optimal, let $p \geq 6$, $\sigma = p \mod 2$, $a = (-2 + 2\sqrt{u \cdot 2^\sigma} + 4u, -u)$, $b = (1 + \frac{1}{2}\sqrt{u \cdot 2^{-\sigma}}, u)$, and $c = (4 - \sqrt{u \cdot 8^\sigma} - 8u, -2u)$. Then one can check that a, b, c are DW numbers, $|c_h| \geq 2|a_h b_h|$, $d_h = 2 - \sigma\sqrt{u/2} - 2u$, $\mathrm{ulp}(d_h) = 2u$, and $e = -u + O(u^{3/2})$, so that $d_\ell \sim -6u$ as $u \to 0$ and, therefore, $|d_\ell|/\mathrm{ulp}(d_h) \sim 3$.

To show that the bound $|\delta| \lesssim 11u^2$ stated above is asymptotically optimal, let $a = (-1 - \sqrt{u \cdot 2^\sigma}, -u + 4u^2)$, $b = (1 + \sqrt{u/2^\sigma}, u - 4u^2)$, and $c = (2 + 4\sqrt{u \cdot 2^\sigma} + 4u, -2u + 6u^2)$. Then, again, one can check that a, b, c are DW numbers, that $|c_h| \geq 2|a_h b_h|$, and that the associated relative error satisfies $|d/(ab + c) - 1| \sim 11u^2$ as $u \to 0$. □

Corollary 1. *Let a, b, c, d be as in Theorem 4. If in addition $k_a \geq k_b = k_c = 1/2$, then*

$$|d_\ell| \leq (2k_a + 2)\mathit{ulp}(d_h)$$

and

$$|\delta| \leq \frac{6 + \alpha_a + 2k_a}{1 - (4k_a + 4)u - 2k_a u^2} u^2.$$

Here, $\alpha_a = 2^{\lceil \log_2(4k_a + 5) \rceil - 1}$ and the constants k_a, $k_a + 1$, $k_a + 5/4$, and $k_a + 2$ are assumed to be in $\mathbb{F}$.

Furthermore, the bound on $|d_\ell|$ is asymptotically optimal.

Proof. For the relative error, it suffices to set $k_b = k_c = 1/2$ in the expressions of B and B' of Theorem 4, and to notice that $B < B'$. For the bound on $|d_\ell|$, this would give only $|d_\ell| \leq (2k_a + \frac{5}{2})\mathrm{ulp}(d_h)$. However, this bound can be replaced by $(2k_a + 2)\mathrm{ulp}(d_h)$ by bounding the intermediate quantity $|a_\ell b_h + g|$ by $\varphi(k_a, |a_h|)$, where

$$\varphi(k, x) := (4kx^{-1} + x(1 + u_1))u + 3u(1 + u_1),$$

and deducing that $|d_\ell| \leq (4k_a + 4)u \leq (2k_a + 2)\mathrm{ulp}(d_h)$.

Tightness of the bound on $|d_\ell|$. One can take $(a_\ell, b_\ell, c_\ell) = (-2k_a u, u, 2u)$ and reuse the high parts already used in the proof of Theorem 3 as follows: keep c_h, use b_h to define a_h, and use a to define b_h. Then one can check that $\mathrm{ufp}(d_h) = 2u$ and that $d_\ell \sim (4k_a + 4)u$ for k_a fixed and as $u \to 0$. □

IV. Case study

We illustrate the gap between a classical DW Horner implementation based on the simplified DWTimesDW1 and SloppyDWPlusDW (cf. Sec. I) and FastFMA_DW on a concrete case: the accurate path of CORE-MATH's correctly-rounded binary64 exponential [7]. We highlight (i) the reduced argument and its interval, (ii) the DW evaluation of the polynomial, (iii) the operation counts per Horner step, contrasting them with an extended-FMA formulation.

Approximation polynomial: The reduced argument lies in the interval $\left[-\frac{\ln 2}{2^{13}}, +\frac{\ln 2}{2^{13}}\right]$. It is represented (modulo some tiny rounding errors) by a DW $x = (x_h, x_\ell) = x_h + x_\ell$. An approximation polynomial of degree 6 whose constant coefficient c_0 is forced to be 1 is used for $\frac{e^x - 1}{x}$. Its DW coefficients $(c_{k,h}, c_{k,\ell})$, for $1 \leq k \leq 6$, are given in Table II.

TABLE II
Coefficients of polynomial appearing in the accurate path for exp binary64 implementation of CORE-MATH.

k	$c_{k,h}$ (hex)	$c_{k,\ell}$ (hex)
1	0x1p-1	0x1.712f72ecec2cfp-99
2	0x1.5555555555555p-3	0x1.5555555554d07p-57
3	0x1.5555555555555p-5	0x1.55194d28275dap-59
4	0x1.1111111111111p-7	0x1.12faa0e1c0f7bp-63
5	0x1.6c16c16da6973p-10	-0x1.4ba45ab25d2a3p-64
6	0x1.a01a019eb7f31p-13	-0x1.9091d845ecd36p-67

DW Horner evaluation in CORE-MATH: Each Horner step is implemented by a simplified (no renormalization as explained in Sec. I) DWTIMESDW1 followed by SLOPPYDWPLUSDW. This is in line with our motivation, since $|x|$ is tiny and the coefficients decrease.

Experimental comparison: For the degree-6 polynomial, a Horner scheme using FASTFMA_DW takes $6 \times (\mathbf{4\ fma\ +\ 2\ add/sub})$, while a CORE-MATH DW-Horner takes $6 \times (\mathbf{3\ mul\ +\ 1\ fma\ +\ 7\ add/sub})$. This operation gain can be confirmed experimentally. We compare the CORE-MATH *original* implementation to a *modified* version in which the polynomial evaluation considered uses FASTFMA_DW. To ensure that the accurate path is always taken, we select inputs from the hard-to-round test set `exp.wc` shipped with CORE-MATH and retain the 106,471 cases whose binary exponent lies between -4 and 8. These exponents avoid underflow/overflow while keeping a sufficiently large sample. We adapted the performance framework provided with CORE-MATH to measure the execution time on this restricted set of hard cases. This also allows us to show experimentally that all hard-to-round cases pass with these modifications, indicating that rounding errors remain small along this path.

We benchmark on an Intel Core i5-1145G7 (2.60 GHz) with `gcc` 11.4.0/`glibc` 2.35 using `RDTSC` counter (natively available on this x86-64 platform); values are means over 20 trials, in cycles per call.

Table III reports results on hard-to-round inputs. Our modified `exp` reduces the average latency from 83.5 to 73.7 cycles per call (a 11.7% improvement) and the reciprocal throughput from 53.5 to 47.4 cycles per call (a 11.4% improvement). For reference, `glibc` 2.35 achieves 24.6 and 8.9 cycles per call, respectively, on the same benchmark.

As a reference, we also report measurements obtained with the default inputs generated by the CORE-MATH framework (random inputs in $[-10, 10]$), which illustrate that the accurate path is substantially more expensive than the *average* one.

TABLE III
PERFORMANCE OF BINARY64 EXP (CYCLES PER CALL) ON HARD-TO-ROUND INPUTS (EXP.WC, EXPONENT IN $[-4, 8]$) AND ON DEFAULT RANDOM INPUTS ($[-10, 10]$).

CORE-MATH, accurate path (`exp.wc`)	Latency	Recip. thr.
`exp` (original)	83.5	53.5
`exp` (FASTFMA_DW)	73.7	47.4
`glibc` 2.35 `exp`	24.6	8.9
Default random inputs ($[-10, 10]$)	**Latency**	**Recip. thr.**
CORE-MATH `exp`	37.1	12.2
`glibc` 2.35 `exp`	24.1	8.8

V. DISCUSSION

We report results on the *accurate path* because, in a scalar implementation of `exp`, the *fast path* can often be implemented using only binary64 arithmetic thanks to the use of lookup tables (range reduction and/or tabulated constants), and scalar code can typically afford larger tables without the same constraints as SIMD code. In that setting, the fast path is already tuned around double-only operations, so the most natural place to assess the benefit of our DW FMA kernels was the accurate path. These experiments thus provide evidence of a speedup in the accurate path, but it remains to evaluate the impact of our kernels on other fast-path designs, especially on vector implementations where the fast-path often contains DW Horner evaluations.

Concerning the opposite regime, where $|c_h| < 2|a_h b_h|$, very different situations can occur. For example, if $|a_h b_h| \le |c_h| < 2|a_h b_h|$ and $a_h b_h$ and c_h have the same sign, then cancellation is avoided and [14, Lem. 7] shows further that $c_h - d_h \in \mathbb{F}$, so the analysis of the previous sections can be adapted. On the other hand, if $|a_h b_h| \approx |c_h|$ and $a_h b_h c_h < 0$, then FASTFMA_DW may of course suffer heavy cancellation and produce a totally wrong result. For example, with $a_h = 1$, $a_\ell = -u/4$, $b_h = 1$, $b_\ell = u/2$, $c_h = -1$, $c_\ell = -u/4$, this algorithm returns $d_h + d_\ell = 0$, while the exact value is $-u^2/8$. This motivates either guarding FASTFMA_DW (detecting potential cancellation and switching kernels) or using a more robust formulation in this general subcase.

REFERENCES

[1] Y. Hida, X. S. Li, and D. H. Bailey, "Algorithms for quad-double precision floating-point arithmetic," in *ARITH Proc.*, 2001.

[2] M. Joldeş, J.-M. Muller, and V. Popescu, "Tight and rigorous error bounds for basic building blocks of double-word arithmetic," *ACM Trans. Math. Software*, 2017.

[3] J.-M. Muller, N. Brunie, F. de Dinechin, C.-P. Jeannerod, M. Joldeş, V. Lefèvre, G. Melquiond, N. Revol, and S. Torres, *Handbook of Floating-Point Arithmetic*. Birkhäuser, 2018.

[4] S. Boldo and J.-M. Muller, "Exact and approximated error of the FMA," *IEEE Trans. Comp.*, 2011.

[5] S. M. Rump and M. Lange, "Faithfully rounded floating-point computation," *ACM Trans. Math. Soft.*, 2020.

[6] M. Cornea, J. Harrison, and P. T. P. Tang, *Scientific Computing on Itanium®-based Systems*. Intel Press, 2002.

[7] A. Sibidanov, P. Zimmermann, and S. Glondu, "The CORE-MATH project," in *ARITH Proc.*, 2022.

[8] N. Brisebarre, G. Hanrot, J. Muller, and P. Zimmermann, "Correctly-rounded evaluation of a function: Why, how, and at what cost?" *ACM Computing Surveys*, 2026.

[9] N. Brisebarre and S. Chevillard, "Efficient polynomial L^∞-approximations," in *ARITH Proc.*, 2007.

[10] N. Brisebarre and G. Hanrot, "Floating-point L^2-approximations," in *ARITH Proc.*, 2007.

[11] S. Chevillard, M. Joldeş, and C. Lauter, "Sollya: An environment for the development of numerical codes," in *ICMS Proc.*, 2010.

[12] T. Kouya, "Performance evaluation of an efficient double-double BLAS1 function with error-free transformation and its application to explicit extrapolation methods," in *ARITH Proc.*, 2019.

[13] T. Koizumi, T. Tsumura, H. Irie, and S. Sakai, "A fast implementation of perfect-accuracy double-precision exponential function," 2023, IPSJ SIG Tech. Rep., 2023-HPC-192 (https://ipsj.ixsq.nii.ac.jp/records/231099).

[14] K. Ozaki and T. Koizumi, "Fast and accurate algorithm for matrix multiplication using fused multiply-add," *JSIAM Letters*, 2025.

[15] ——, "Fast and accurate algorithms for matrix multiplication using fused multiply-add and their rounding error analysis," 2025, preprint.

[16] S. M. Rump, T. Ogita, and S. Oishi, "Accurate floating-point summation part I: Faithful rounding," *SIAM J. Sci. Comput.*, 2008.

[17] T. Hubrecht, C.-P. Jeannerod, and J.-M. Muller, "Useful applications of correctly-rounded operators of the form $ab + cd + e$," in *ARITH Proc.*, 2024.