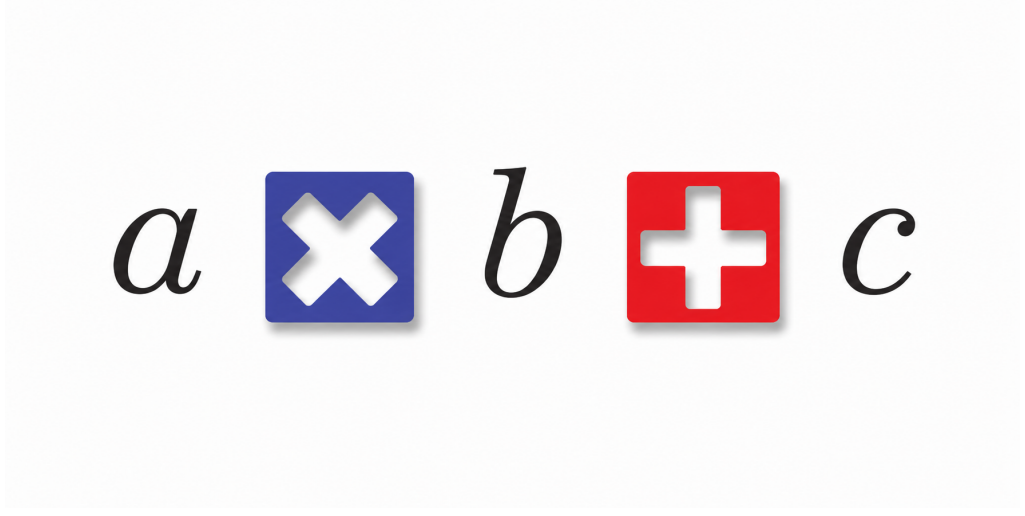
Extended-Precision FMA under Parameterized Double-Word Overlap

Tight error bounds and examples

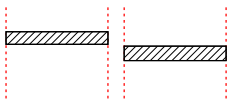
Claude-Pierre Jeannerod*, Mioara Joldes**, Nicolas Louvet*, Jean-Michel Muller*

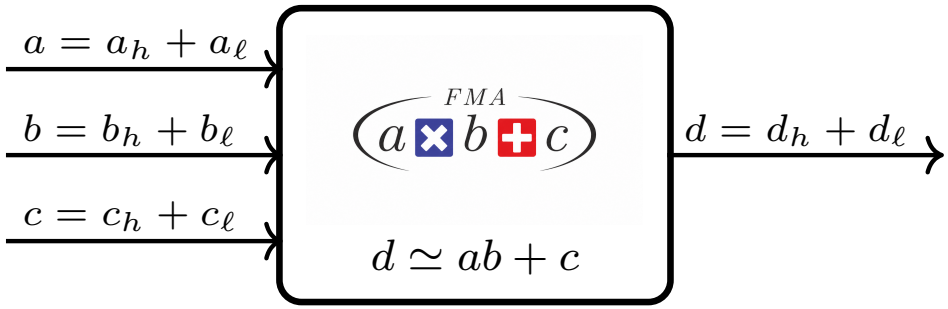


33rd IEEE International Symposium on Computer Arithmetic
ARITH 2026

*Inria, Université Lyon 1, CNRS, ENSL, LIP, Lyon, France; **CNRS, LAAS, Toulouse, France

Goal: Fast **Double-Word** Multiply-Add

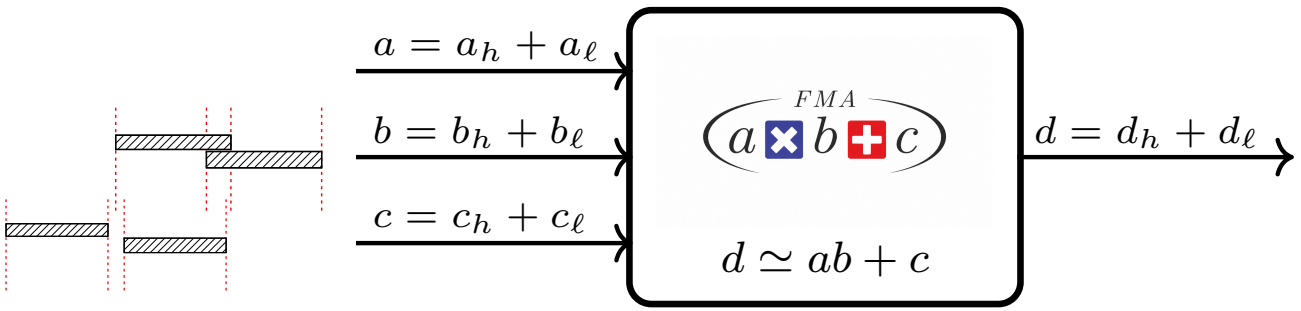




DW inputs with
parameterized overlap

DW output with overlap
and worst-case tight error bounds

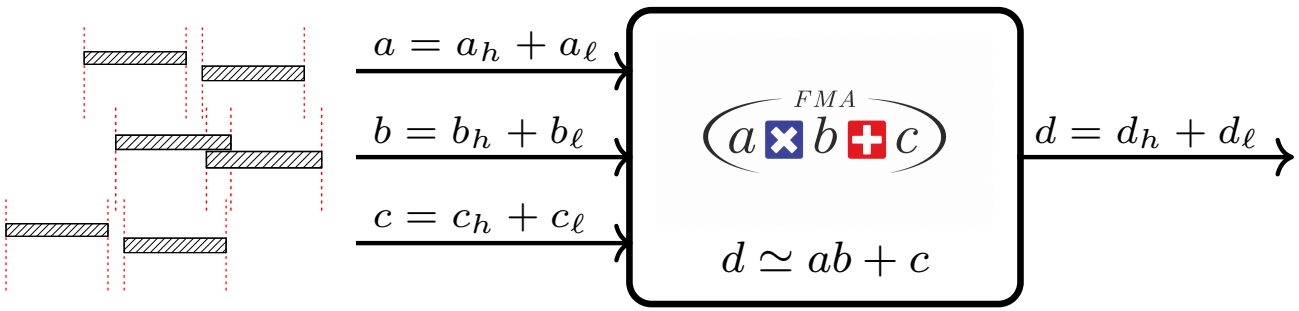
Goal: Fast **Double-Word** Multiply-Add



DW inputs with
parameterized overlap

DW output with overlap
and worst-case tight error bounds

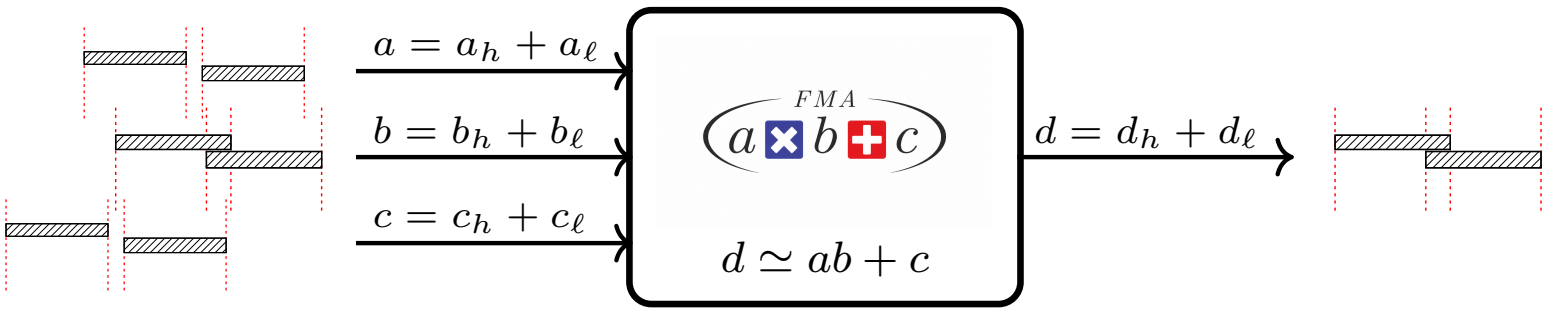
Goal: Fast **Double-Word** Multiply-Add



DW inputs with
parameterized overlap

DW output with overlap
and worst-case tight error bounds

Goal: Fast **Double-Word** Multiply-Add



DW inputs with
parameterized overlap

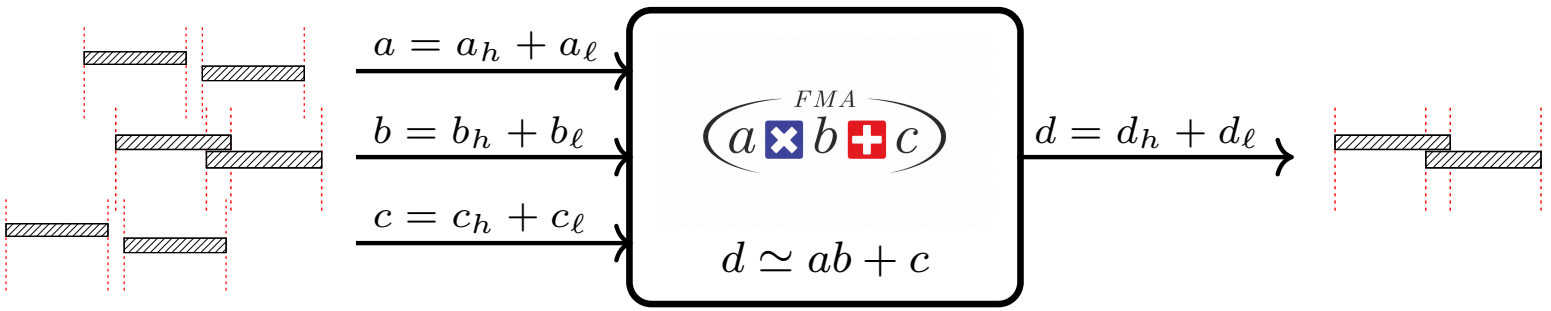
DW output with overlap
and worst-case tight error bounds

Error bounds:

$$\begin{cases} \text{relative error of } d_h + d_\ell, \\ \text{output overlap of } (d_h, d_\ell), \end{cases}$$

depending on precision and input overlap.

Goal: Fast **Double-Word** Multiply-Add



DW inputs with
parameterized overlap

DW output with overlap
and worst-case tight error bounds

Error bounds:

$$\begin{cases} \text{relative error of } d_h + d_\ell, \\ \text{output overlap of } (d_h, d_\ell), \end{cases}$$

depending on precision and input overlap.

Questions:

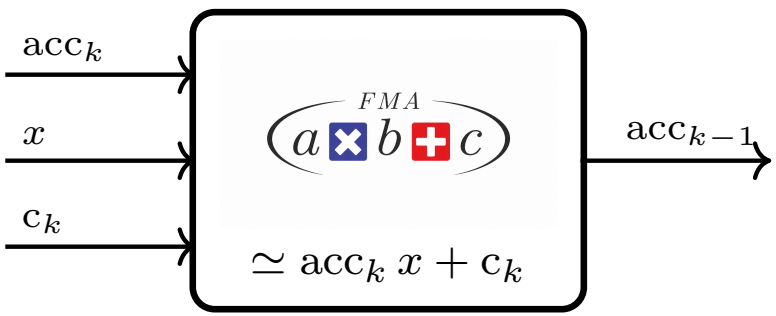
Accuracy	:	how does input overlap affect the error bounds?
Efficiency	:	any improvement over generic DW arithmetic?
Worst cases	:	are the bounds sharp, and are they attainable?

Motivation: a recurring kernel used in **libms**

Example: CORE-MATH accurate path for **binary64** $\exp$

Horner evaluation for a degree-6 polynomial P for $x \in \left[-\frac{\log 2}{2^{13}}, \frac{\log 2}{2^{13}}\right]$.

$$P(x) = c_0 + x\underbrace{(c_1 + x\,(\ldots))}_{\mathrm{acc}}$$

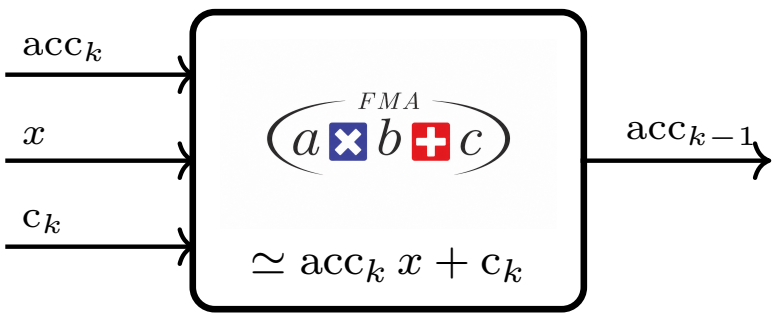


Motivation: a recurring kernel used in **libms**

Example: CORE-MATH accurate path for **binary64** exp

Horner evaluation for a degree-6 polynomial P for $x \in \left[-\frac{\log 2}{2^{13}}, \frac{\log 2}{2^{13}}\right]$.

$$P(x) = c_0 + x\underbrace{(c_1 + x\,(\ldots))}_{\mathrm{acc}}$$



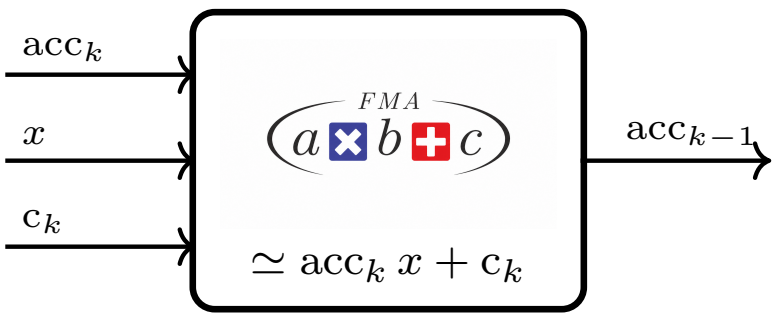
k	$c_{k,h}$ **(hex)**	$c_{k,\ell}$ **(hex)**
1	0x1p-1	0x1.712f72ecec2cfp-99
2	0x1.5555555555555p-3	0x1.5555555554d07p-57
3	0x1.5555555555555p-5	0x1.55194d28275dap-59
4	0x1.1111111111111p-7	0x1.12faa0e1c0f7bp-63
5	0x1.6c16c16da6973p-10	-0x1.4ba45ab25d2a3p-64
6	0x1.a01a019eb7f31p-13	-0x1.9091d845ecd36p-67

Motivation: a recurring kernel used in **libms**

Example: CORE-MATH accurate path for **binary64** exp

Horner evaluation for a degree-6 polynomial P for $x \in \left[-\frac{\log 2}{2^{13}}, \frac{\log 2}{2^{13}}\right]$.

$$P(x) = c_0 + x\underbrace{(c_1 + x\,(\ldots))}_{\mathrm{acc}}$$



Classical DW-Horner step:

DW multiplication + DW addition

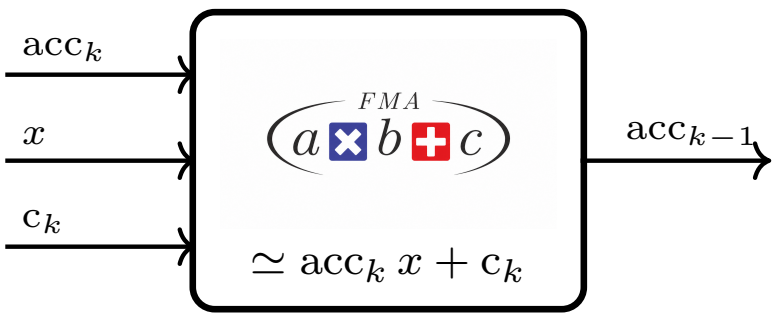
k	$c_{k,h}$ **(hex)**	$c_{k,\ell}$ **(hex)**
1	0x1p-1	0x1.712f72ecec2cfp-99
2	0x1.5555555555555p-3	0x1.5555555554d07p-57
3	0x1.5555555555555p-5	0x1.55194d28275dap-59
4	0x1.1111111111111p-7	0x1.12faa0e1c0f7bp-63
5	0x1.6c16c16da6973p-10	-0x1.4ba45ab25d2a3p-64
6	0x1.a01a019eb7f31p-13	-0x1.9091d845ecd36p-67

Motivation: a recurring kernel used in **libms**

Example: CORE-MATH accurate path for **binary64** exp

Horner evaluation for a degree-6 polynomial P for $x \in \left[-\frac{\log 2}{2^{13}}, \frac{\log 2}{2^{13}}\right]$.

$$P(x) = c_0 + x\underbrace{(c_1 + x\,(\ldots))}_{\mathrm{acc}}$$

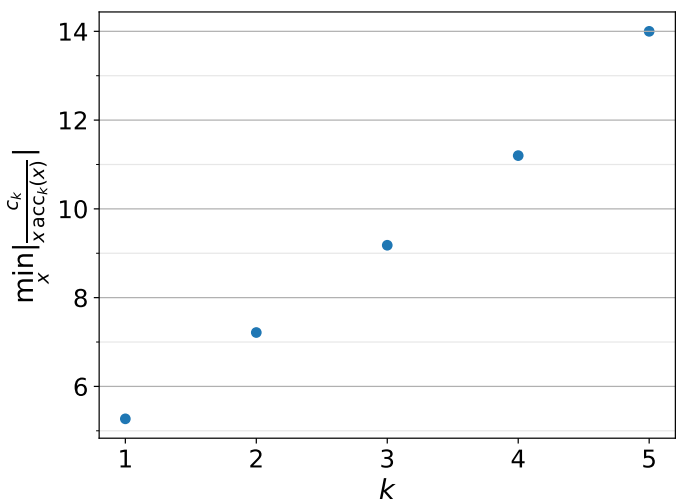


Classical DW-Horner step:

DW multiplication + DW addition

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

Minimum ratio between c_k and $x\,\mathrm{acc}_k$

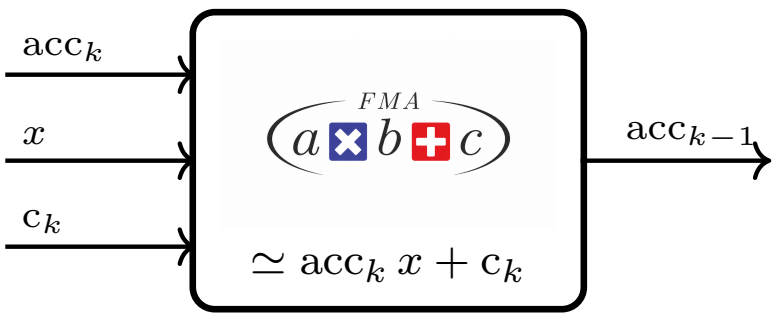


Motivation: a recurring kernel used in **libms**

Example: CORE-MATH accurate path for **binary64** exp

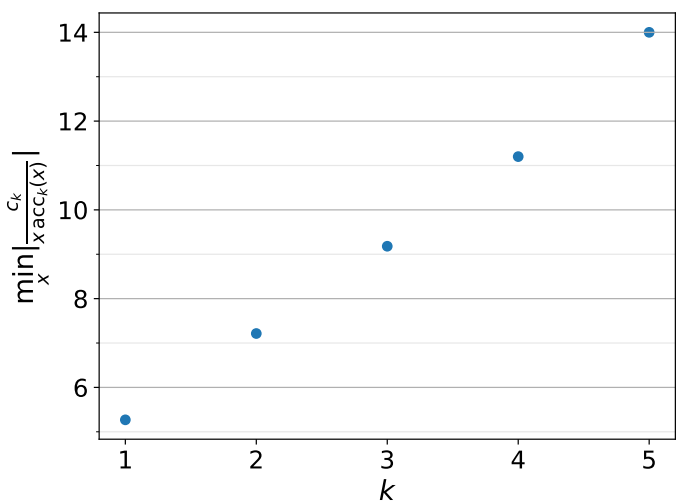
Horner evaluation for a degree-6 polynomial P for $x \in \left[-\frac{\log 2}{2^{13}}, \frac{\log 2}{2^{13}}\right]$.

$$P(x) = c_0 + x\underbrace{(c_1 + x\,(\ldots))}_{\mathrm{acc}}$$



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

Minimum ratio between c_k and $x\,\mathrm{acc}_k$



Classical DW-Horner step:

DW multiplication + DW addition

CORE-MATH optim:
3 mul + 1 fma + 7 add/sub (**binary64**)

Model: floating-point and double-word arithmetic

Floating-point system $\mathbb{F}$, binary, precision p, unbounded exponent range

- Round-to-nearest $\mathrm{RN} : \mathbb{R} \to \mathbb{F}$
- FMA = one rounding: $\mathrm{RN}(ab + c)$
- Useful units: $u = 2^{-p}$
  for $x \in \mathbb{R}\backslash\{0\}$, with $2^e \leq |x| < 2^{e+1}$, define $\mathrm{ulp}(x) = 2^{e-p+1} = 2u\,2^e$
- Error bound:

$$|\mathrm{RN}(x) - x| \leq \frac{1}{2}\,\mathrm{ulp}(x) \leq \frac{1}{2}\,\mathrm{ulp}(\mathrm{RN}(x))$$

Model: floating-point and double-word arithmetic

Floating-point system $\mathbb{F}$, binary, precision p, unbounded exponent range

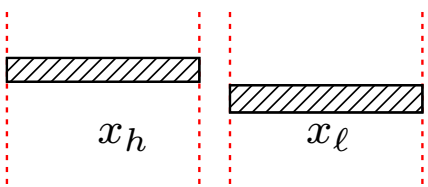
- Round-to-nearest $\mathrm{RN} : \mathbb{R} \to \mathbb{F}$
- FMA = one rounding: $\mathrm{RN}(ab + c)$
- Useful units: $u = 2^{-p}$
  for $x \in \mathbb{R}\backslash\{0\}$, with $2^e \leq |x| < 2^{e+1}$, define $\mathrm{ulp}(x) = 2^{e-p+1} = 2u\,2^e$
- Error bound:

$$|\mathrm{RN}(x) - x| \leq \frac{1}{2}\,\mathrm{ulp}(x) \leq \frac{1}{2}\,\mathrm{ulp}(\mathrm{RN}(x))$$

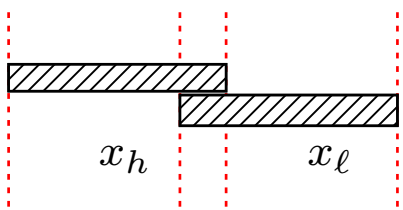
Double-word representation [Dekker'71, Priest'91, Bailey et al.'01, LangeRump'20]

- $x = (x_h, x_\ell) \in \mathbb{F}^2$ identified with the unevaluated sum $x = x_h + x_\ell$

Classical case: $x_h = \mathrm{RN}(x_h + x_\ell)$,
$|x_\ell| \leq \frac{1}{2}\,\mathrm{ulp}(x_h)$



Parametrized overlap:
$|x_\ell| \leq k_x\ \mathrm{ulp}(x_h), \quad k_x > 0$



FastTwoFMA[OzakiKoizumi'25]: $d_h + d_\ell \simeq ab + c$

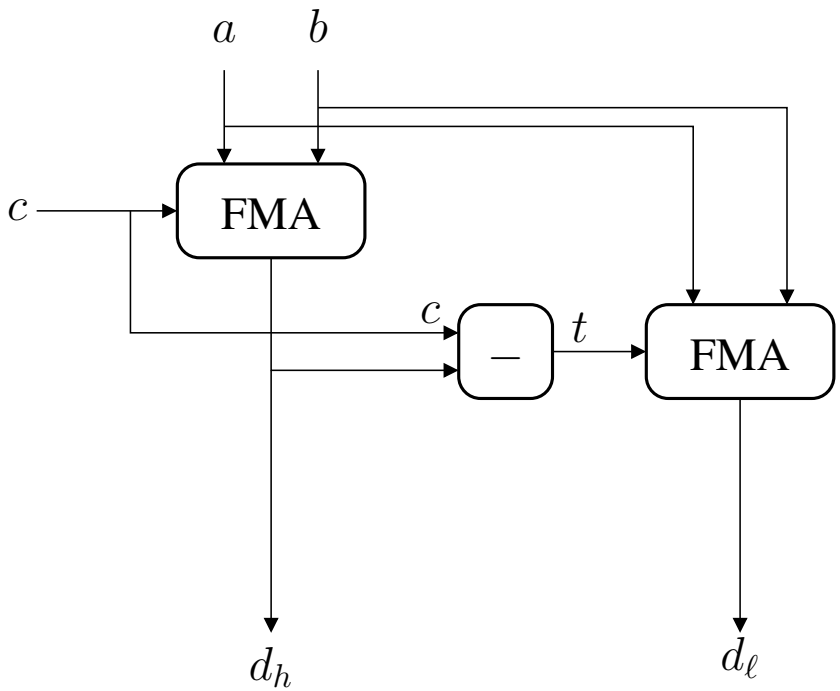
Algorithm FastTwoFMA$(a, b, c \in \mathbb{F})$

$d_h \leftarrow \mathrm{RN}(ab + c)$

$t \leftarrow \mathrm{RN}(c - d_h)$

$d_\ell \leftarrow \mathrm{RN}(ab + t)$

return (d_h, d_ℓ)



FastTwoFMA[OzakiKoizumi'25]: $d_h + d_\ell \simeq ab + c$

Algorithm FastTwoFMA$(a, b, c \in \mathbb{F})$

$d_h \leftarrow \mathrm{RN}(ab + c)$

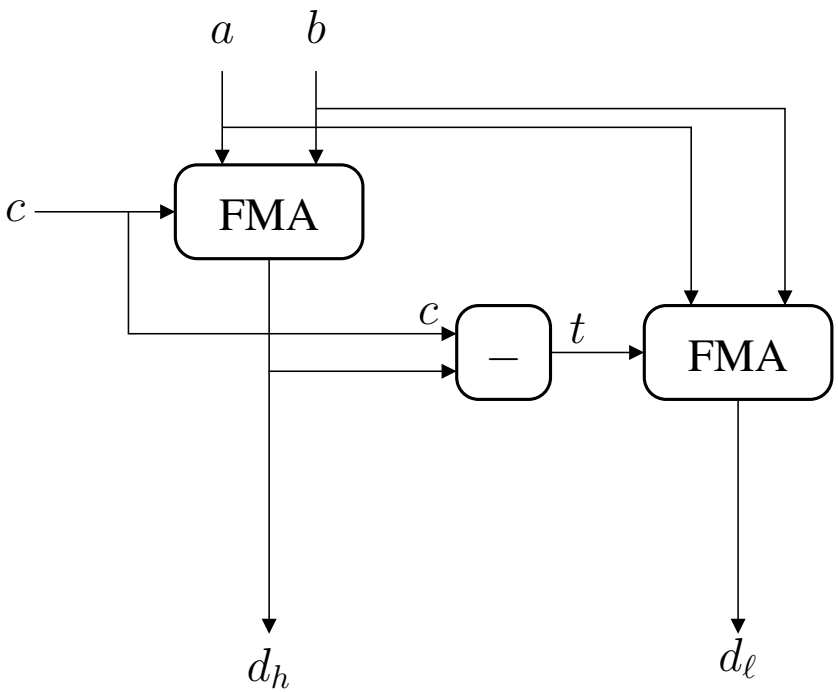
$t \leftarrow \mathrm{RN}(c - d_h)$

$d_\ell \leftarrow \mathrm{RN}(ab + t)$

return (d_h, d_ℓ)

Dominance condition $|c| \geq 2|ab|$:

$$\Rightarrow c - d_h \in \mathbb{F}$$



FastTwoFMA[OzakiKoizumi'25]: $d_h + d_\ell \simeq ab + c$

Algorithm FastTwoFMA$(a, b, c \in \mathbb{F})$

$d_h \leftarrow \mathrm{RN}(ab + c)$

$t \leftarrow \mathrm{RN}(c - d_h)$

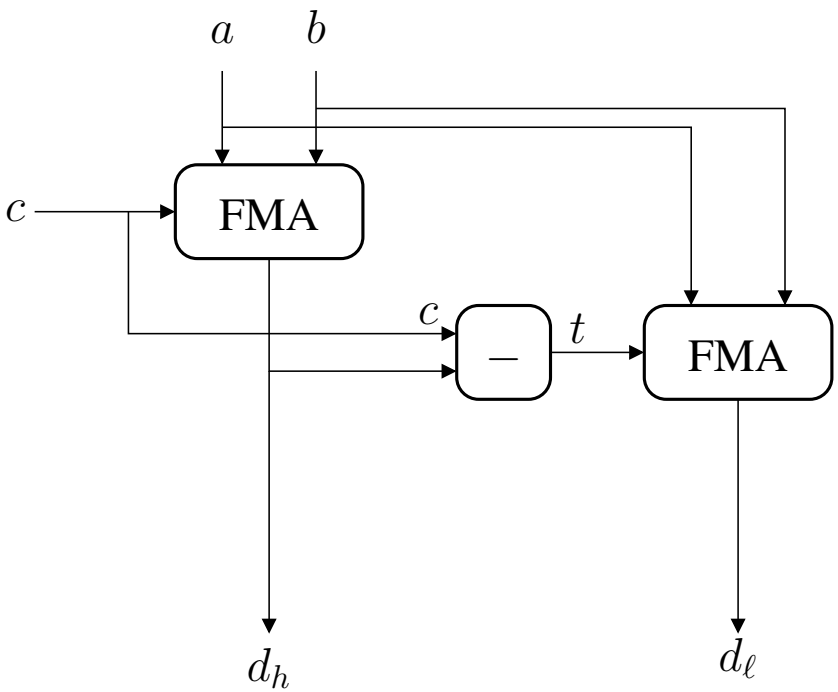
$d_\ell \leftarrow \mathrm{RN}(ab + t)$

return (d_h, d_ℓ)

Dominance condition $|c| \geq 2|ab|$:

$\Rightarrow c - d_h \in \mathbb{F}$

$\Rightarrow d_\ell = \mathrm{RN}(ab + c - d_h)$



FastTwoFMA[OzakiKoizumi'25]: $d_h + d_\ell \simeq ab + c$

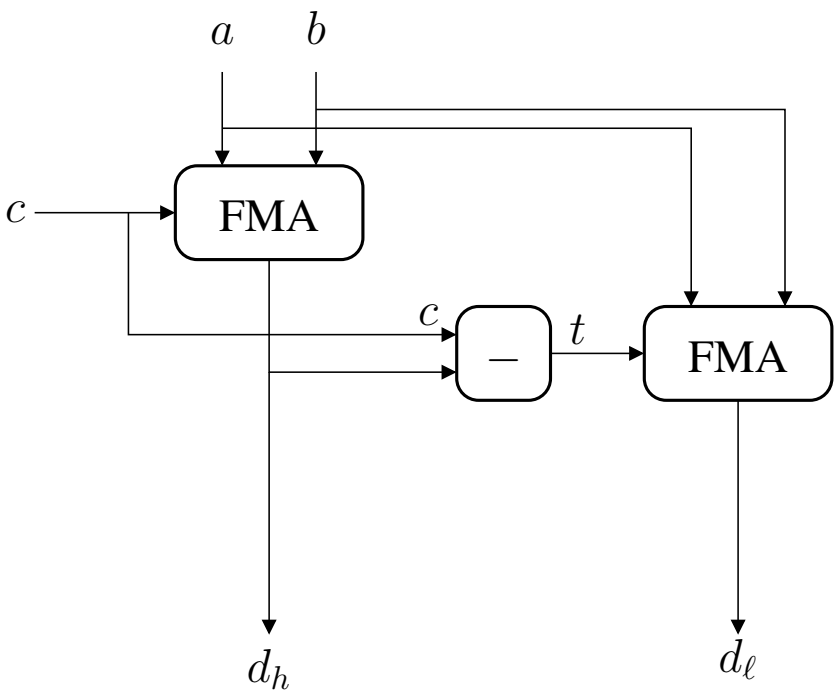
Algorithm FastTwoFMA$(a, b, c \in \mathbb{F})$

$d_h \leftarrow \mathrm{RN}(ab + c)$

$t \leftarrow \mathrm{RN}(c - d_h)$

$d_\ell \leftarrow \mathrm{RN}(ab + t)$

return (d_h, d_ℓ)



Dominance condition $|c| \geq 2|ab|$:

$\Rightarrow c - d_h \in \mathbb{F}$

$\Rightarrow d_\ell = \mathrm{RN}(ab + c - d_h)$

$d = d_h + d_\ell$

FastTwoFMA[OzakiKoizumi'25]: $d_h + d_\ell \simeq ab + c$

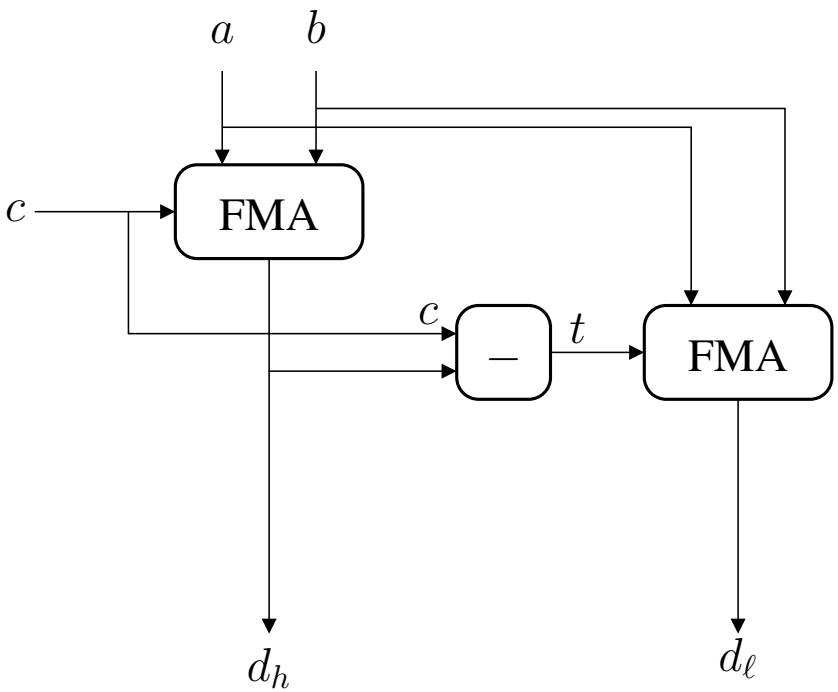
Algorithm FastTwoFMA$(a, b, c \in \mathbb{F})$

$d_h \leftarrow \mathrm{RN}(ab + c)$

$t \leftarrow \mathrm{RN}(c - d_h)$

$d_\ell \leftarrow \mathrm{RN}(ab + t)$

return (d_h, d_ℓ)



Dominance condition $|c| \geq 2|ab|$:

$\Rightarrow c - d_h \in \mathbb{F}$

$\Rightarrow d_\ell = \mathrm{RN}(ab + c - d_h)$

$d = d_h + d_\ell$

- $|(ab + c) - d| = |\underbrace{(ab + c) - d_h}_{r} - d_\ell|$

  $= |r - \mathrm{RN}(r)|$

FastTwoFMA[OzakiKoizumi'25]: $d_h + d_\ell \simeq ab + c$

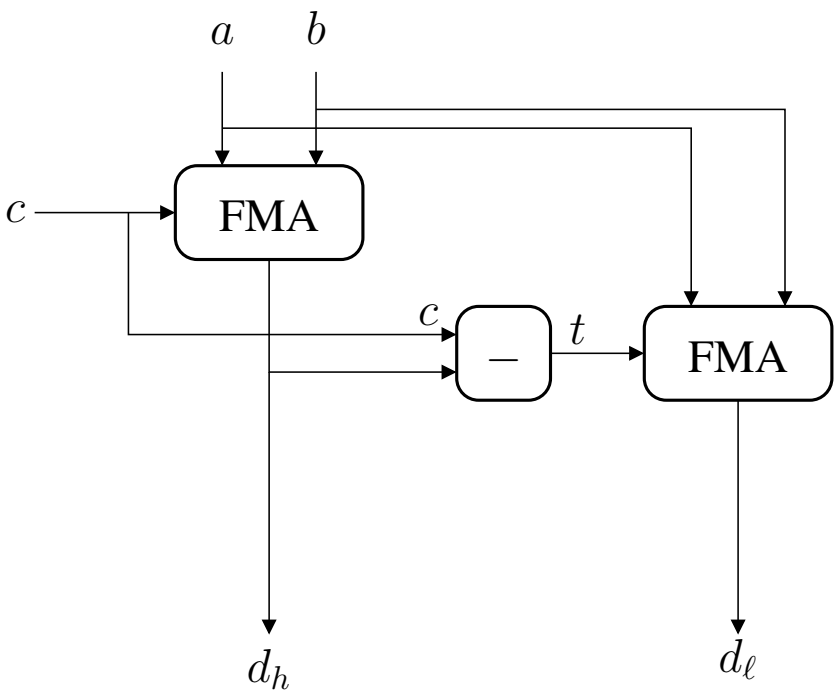
Algorithm FastTwoFMA$(a, b, c \in \mathbb{F})$

$d_h \leftarrow \mathrm{RN}(ab + c)$

$t \leftarrow \mathrm{RN}(c - d_h)$

$d_\ell \leftarrow \mathrm{RN}(ab + t)$

return (d_h, d_ℓ)



Dominance condition $|c| \geq 2|ab|$:

$\Rightarrow c - d_h \in \mathbb{F}$

$\Rightarrow d_\ell = \mathrm{RN}(ab + c - d_h)$

$d = d_h + d_\ell$

- $|(ab + c) - d| = |\underbrace{(ab + c) - d_h}_{r} - d_\ell|$

  $= |r - \mathrm{RN}(r)|$

- $|\underbrace{ab + c - d_h}_{r}| \leq \frac{1}{2}\,\mathrm{ulp}(ab + c)$

  $\Rightarrow |\,\mathrm{RN}(r)| \leq \frac{1}{2}\,\mathrm{ulp}(ab + c)$

FastTwoFMA[OzakiKoizumi'25]: $d_h + d_\ell \simeq ab + c$

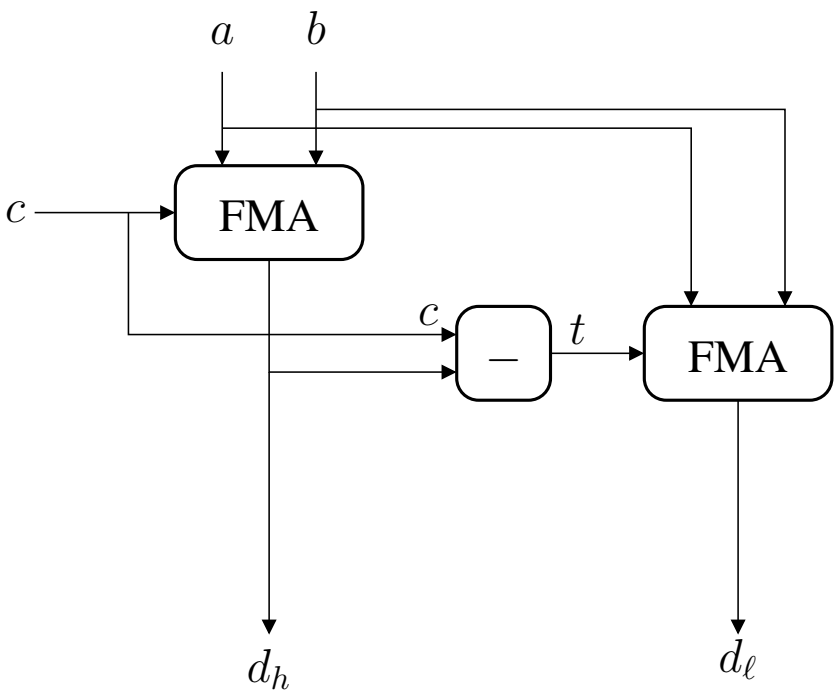
Algorithm FastTwoFMA$(a, b, c \in \mathbb{F})$

$d_h \leftarrow \mathrm{RN}(ab + c)$

$t \leftarrow \mathrm{RN}(c - d_h)$

$d_\ell \leftarrow \mathrm{RN}(ab + t)$

return (d_h, d_ℓ)



Dominance condition $|c| \geq 2|ab|$:

$\Rightarrow c - d_h \in \mathbb{F}$

$\Rightarrow d_\ell = \mathrm{RN}(ab + c - d_h)$

$d = d_h + d_\ell$

- $|(ab + c) - d| = |\underbrace{(ab + c) - d_h}_{r} - d_\ell|$
  $= |r - \mathrm{RN}(r)| < \frac{1}{2}u^2\,|ab + c|$
- $|\underbrace{ab + c - d_h}_{r}| \leq \frac{1}{2}\,\mathrm{ulp}(ab + c)$
  $\Rightarrow |\,\mathrm{RN}(r)| \leq \frac{1}{2}\,\mathrm{ulp}(ab + c)$

FastTwoFMA[OzakiKoizumi'25]: $d_h + d_\ell \simeq ab + c$

Algorithm FastTwoFMA$(a, b, c \in \mathbb{F})$

$d_h \leftarrow \mathrm{RN}(ab + c)$

$t \leftarrow \mathrm{RN}(c - d_h)$

$d_\ell \leftarrow \mathrm{RN}(ab + t)$

return (d_h, d_ℓ)

Dominance condition $|c| \geq 2|ab|$:

$\Rightarrow c - d_h \in \mathbb{F}$

$\Rightarrow d_\ell = \mathrm{RN}(ab + c - d_h)$

$d = d_h + d_\ell$

Theorem (FP inputs):

If $c - d_h \in \mathbb{F}$ then $d = (ab + c)(1 + \delta)$,

$$|\delta| < \frac{1}{2}u^2 \quad \text{and} \quad |d_\ell| \leq \frac{1}{2}\,\mathrm{ulp}(d_h).$$

The bounds are (asymptotically) tight.

Proof:

- $|(ab + c) - d| = |\underbrace{(ab + c) - d_h}_{r} - d_\ell|$

  $= |r - \mathrm{RN}(r)| < \frac{1}{2}u^2\,|ab + c|$

- $|\underbrace{ab + c - d_h}_{r}| \leq \frac{1}{2}\,\mathrm{ulp}(ab + c)$

  $\Rightarrow |\,\mathrm{RN}(r)| \leq \frac{1}{2}\,\mathrm{ulp}(ab + c)$

FastTwoFMA[OzakiKoizumi'25]: $d_h + d_\ell \simeq ab + c$

Algorithm FastTwoFMA$(a, b, c \in \mathbb{F})$

$d_h \leftarrow \mathrm{RN}(ab + c)$

$t \leftarrow \mathrm{RN}(c - d_h)$

$d_\ell \leftarrow \mathrm{RN}(ab + t)$

return (d_h, d_ℓ)

For $(a, b, c) = (u, 1, 1)$

$d_h = \mathrm{RN}(u + 1) = 1$

$c - d_h = 0 \in \mathbb{F}$

$d_\ell = \mathrm{RN}(u + 1 - 1) = u$

$d = d_h + d_\ell$

Theorem (FP inputs):

If $c - d_h \in \mathbb{F}$ then $d = (ab + c)(1 + \delta)$,

$$|\delta| < \frac{1}{2}u^2 \quad \text{and} \quad |d_\ell| \leq \frac{1}{2}\,\mathrm{ulp}(d_h).$$

The bounds are (asymptotically) tight.

Worst cases:

- $(a, b, c) = (u, 1, 1)$

  $\Rightarrow \frac{|d_\ell|}{\mathrm{ulp}(d_h)} = \frac{u}{2u} = \frac{1}{2}$

FastTwoFMA[OzakiKoizumi'25]: $d_h + d_\ell \simeq ab + c$

Algorithm FastTwoFMA$(a, b, c \in \mathbb{F})$

$d_h \leftarrow \mathrm{RN}(ab + c)$

$t \leftarrow \mathrm{RN}(c - d_h)$

$d_\ell \leftarrow \mathrm{RN}(ab + t)$

return (d_h, d_ℓ)

For $(a, b, c) = (1 - u, \frac{3}{2}u, 1)$

$d_h = 1 + 2u$

$c - d_h = -2u \in \mathbb{F}$

$d_\ell = -u/2 - 2u^2$

$d = d_h + d_\ell$

Theorem (FP inputs):

If $c - d_h \in \mathbb{F}$ then $d = (ab + c)(1 + \delta)$,

$$|\delta| < \frac{1}{2}u^2 \quad \text{and} \quad |d_\ell| \le \frac{1}{2}\,\mathrm{ulp}(d_h).$$

The bounds are (asymptotically) tight.

Worst cases:

- $(a, b, c) = (u, 1, 1)$

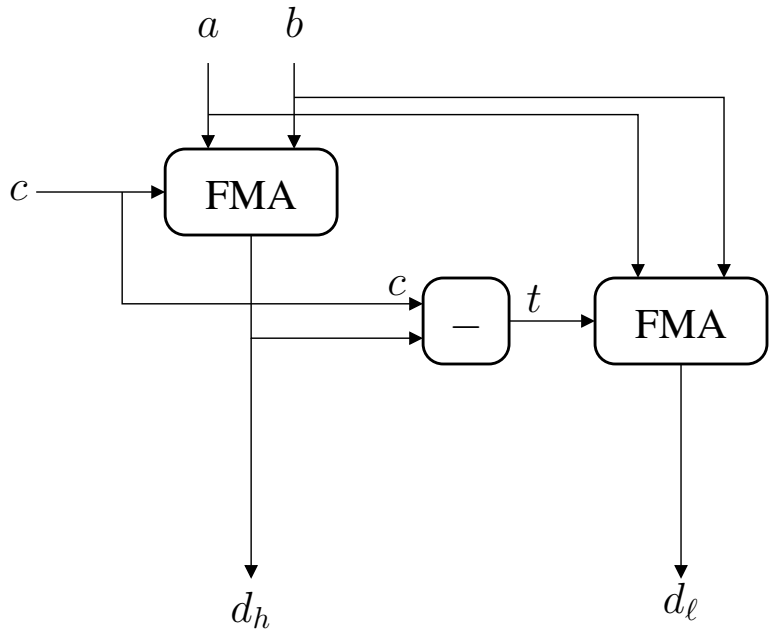
  $\Rightarrow \frac{|d_\ell|}{\mathrm{ulp}(d_h)} = \frac{u}{2u} = \frac{1}{2}$
- $(a, b, c) = (1 - u, \frac{3}{2}u, 1)$

  $\Rightarrow |d - (ab + c)| = \frac{1}{2}u^2$

  $\Rightarrow |\delta| \sim \frac{1}{2}u^2$ as $u \to 0$

From FP inputs to DW inputs

[Ozaki–Koizumi]$\rightsquigarrow$ improved/new error analyses



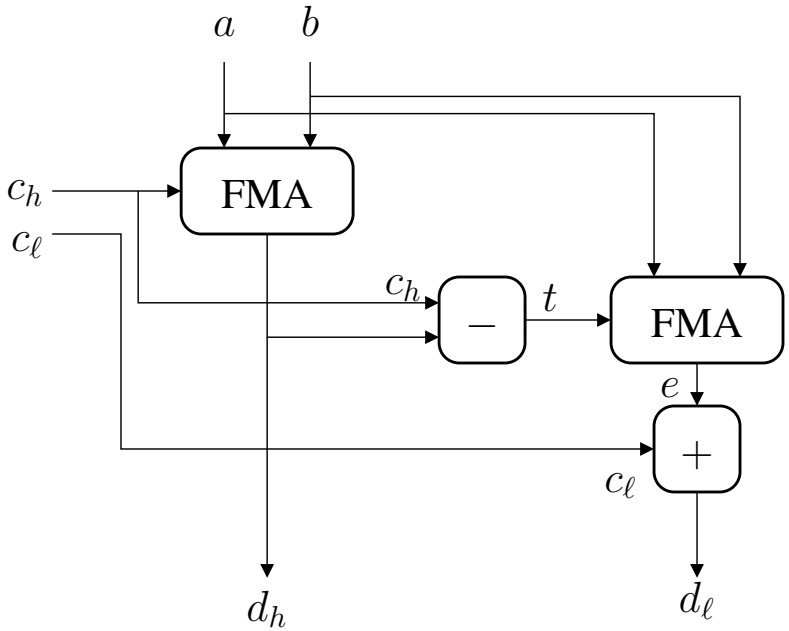
FastTwoFMA : $\mathbb{F} \times \mathbb{F} + \mathbb{F} \to \mathrm{DW}$,

Same structure:

1. high part $\rightsquigarrow$ one FMA;
2. residual $\rightsquigarrow$ one exact subtraction and one FMA;

From FP inputs to DW inputs

[Ozaki–Koizumi]$\rightsquigarrow$ improved/new error analyses



FastTwoFMA : $\mathbb{F} \times \mathbb{F} + \mathbb{F} \to \mathrm{DW}$,

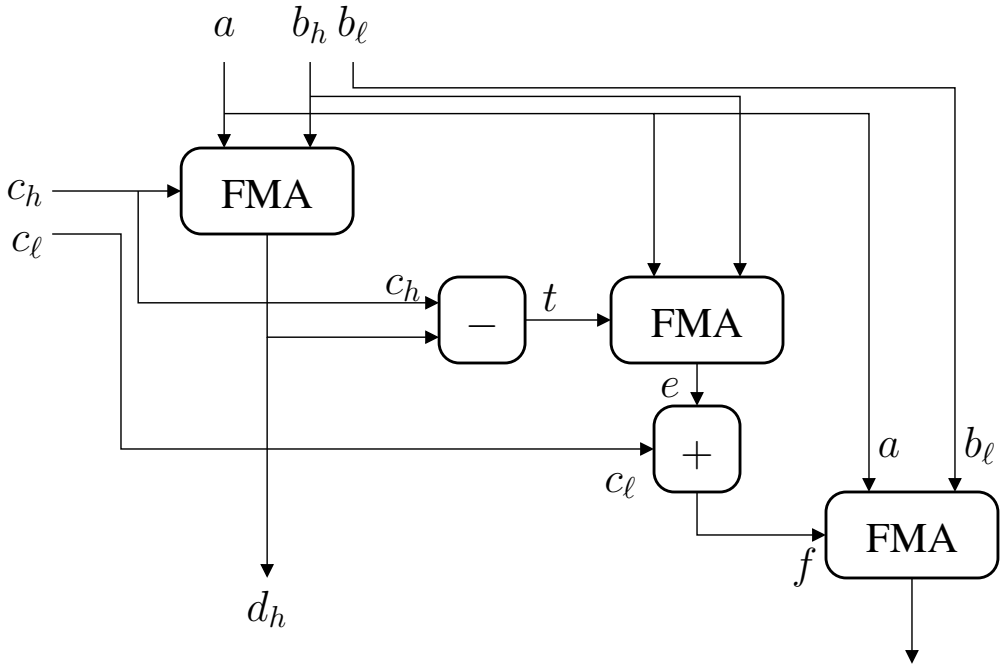
FastTwoFMA_s : $\mathbb{F} \times \mathbb{F} + \mathrm{DW} \to \mathrm{DW}$,

Same structure:

1. high part $\rightsquigarrow$ one FMA;
2. residual $\rightsquigarrow$ one exact subtraction and one FMA;
3. add DW low terms with a rounded operation and FMA.

From FP inputs to DW inputs

[Ozaki–Koizumi]$\rightsquigarrow$ improved/new error analyses



FastTwoFMA : $\mathbb{F} \times \mathbb{F} + \mathbb{F} \to \mathrm{DW}$,

FastTwoFMA_s : $\mathbb{F} \times \mathbb{F} + \mathrm{DW} \to \mathrm{DW}$,

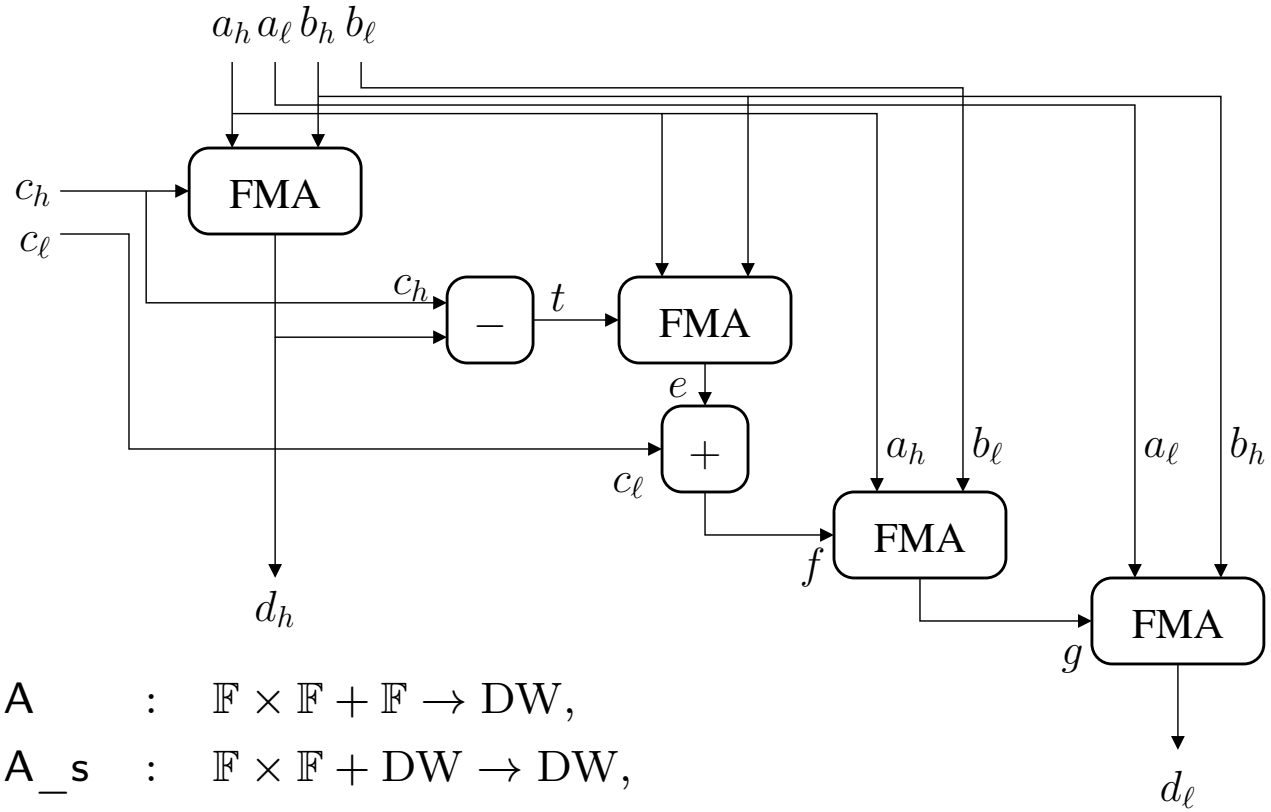
FastFMA_DWh : $\mathbb{F} \times \mathrm{DW} + \mathrm{DW} \to \mathrm{DW}$,

Same structure:

1. high part $\rightsquigarrow$ one FMA;
2. residual $\rightsquigarrow$ one exact subtraction and one FMA;
3. add DW low terms with a rounded operation and FMA.

From FP inputs to DW inputs

[Ozaki–Koizumi]$\rightsquigarrow$ improved/new error analyses



FastTwoFMA : $\mathbb{F} \times \mathbb{F} + \mathbb{F} \to \mathrm{DW}$,

FastTwoFMA_s : $\mathbb{F} \times \mathbb{F} + \mathrm{DW} \to \mathrm{DW}$,

FastFMA_DWh : $\mathbb{F} \times \mathrm{DW} + \mathrm{DW} \to \mathrm{DW}$,

FastFMA_DW : $\mathrm{DW} \times \mathrm{DW} + \mathrm{DW} \to \mathrm{DW}$.

Same structure:

1. high part $\rightsquigarrow$ one FMA;
2. residual $\rightsquigarrow$ one exact subtraction and one FMA;
3. add DW low terms with a rounded operation and FMA.

Main bounds for classical DW overlap

Under the dominance condition

$$|c| \geq 2|ab|,$$

and classical DW input overlap

$$|\__\ell| \leq \frac{1}{2}\,\mathrm{ulp}(\__h),$$

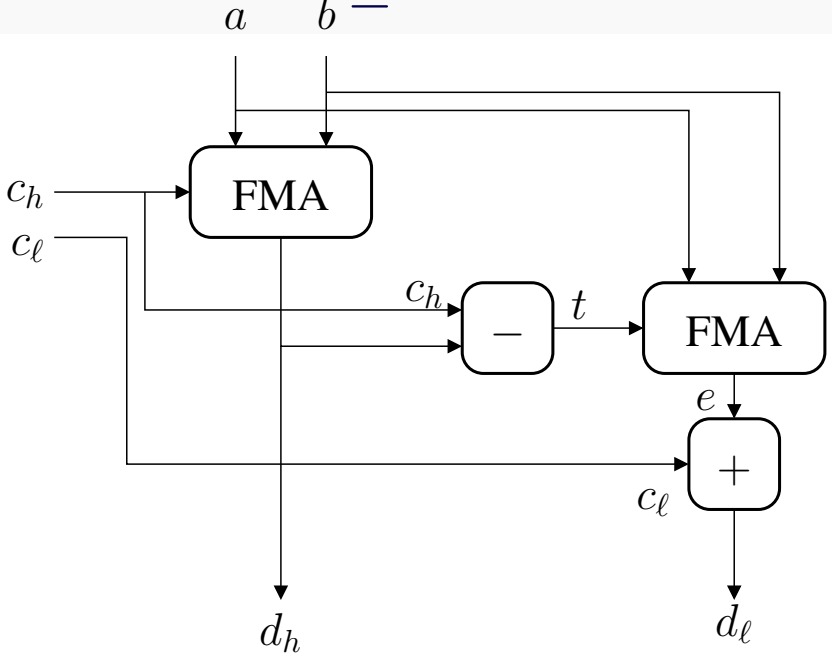
we obtain the following worst-case bounds for $d_h + d_\ell = (ab + c)(1 + \delta)$:

Kernel and signature	bound on $\lvert d_\ell \rvert$	bound on $\lvert \delta \rvert$
FastTwoFMA : $\mathbb{F} \times \mathbb{F} + \mathbb{F} \to \mathrm{DW}$	$\leq \frac{1}{2}\,\mathrm{ulp}(d_h)$	$< \frac{1}{2}u^2$
FastTwoFMA_s : $\mathbb{F} \times \mathbb{F} + \mathrm{DW} \to \mathrm{DW}$	$\leq \frac{3}{2}\,\mathrm{ulp}(d_h)$	$\leq \dfrac{2u^2}{1-2u}$
FastFMA_DWh : $\mathbb{F} \times \mathrm{DW} + \mathrm{DW} \to \mathrm{DW}$	$\leq \frac{5}{2}\,\mathrm{ulp}(d_h)$	$\leq \dfrac{6u^2}{1-4u}$
FastFMA_DW : $\mathrm{DW} \times \mathrm{DW} + \mathrm{DW} \to \mathrm{DW}$	$\leq 3\,\mathrm{ulp}(d_h)$	$\leq \dfrac{11u^2}{1-6u-u^2}$

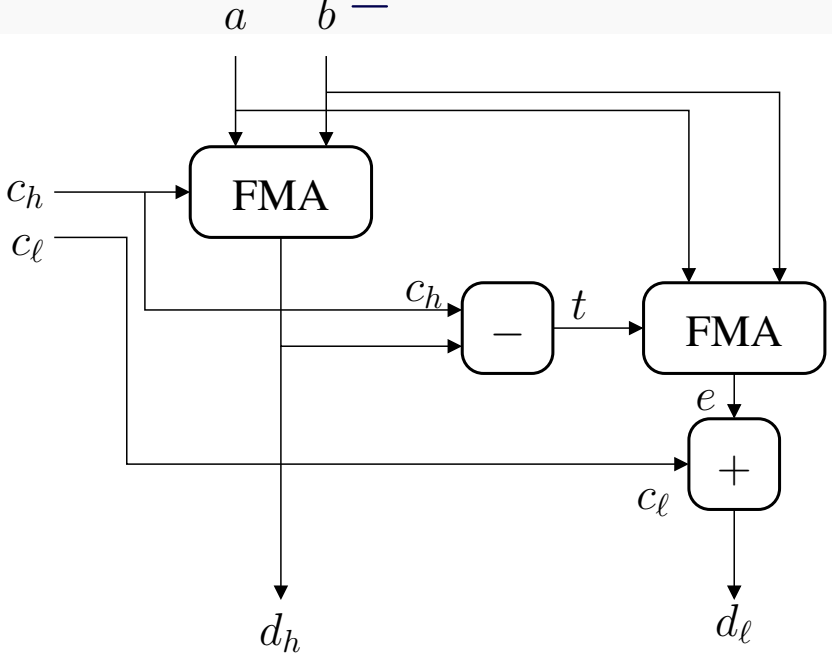
Remark

These bounds are (asymptotically) optimal.

FastTwoFMA_s : $\mathbb{F} \times \mathbb{F} + \mathrm{DW} \to \mathrm{DW}$



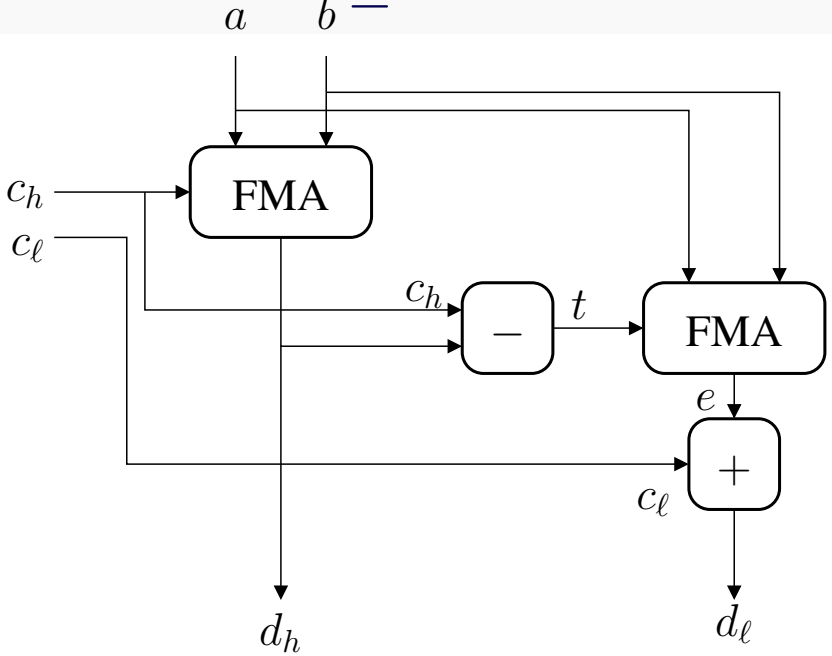
FastTwoFMA_s : $\mathbb{F} \times \mathbb{F} + \mathrm{DW} \to \mathrm{DW}$



If $c_h - d_h \in \mathbb{F}$, then FastTwoFMA gives:

$$e = \mathrm{RN}(ab + c_h - d_h) = \underbrace{ab + c_h} - d_h + \varepsilon_1$$

FastTwoFMA_s : $\mathbb{F} \times \mathbb{F} + \mathrm{DW} \to \mathrm{DW}$

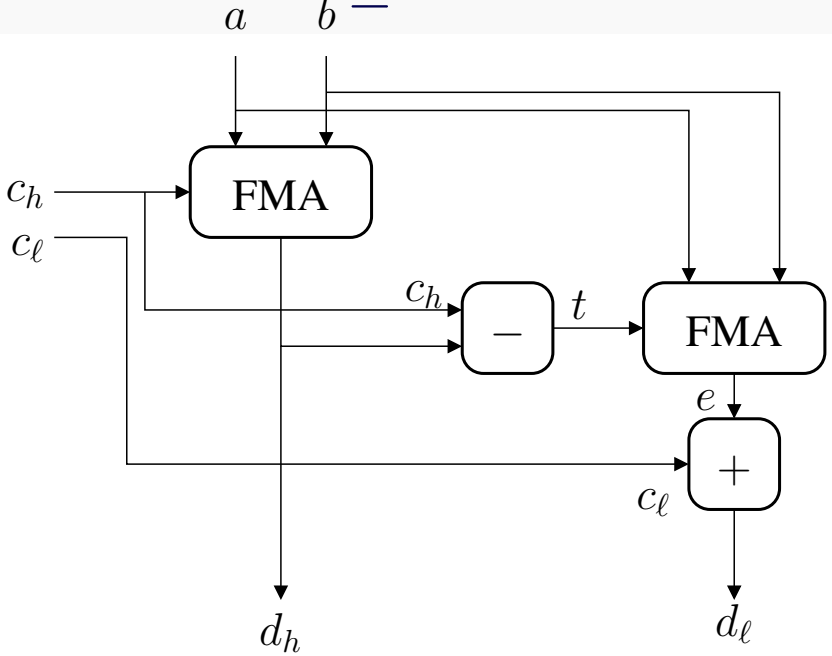


If $c_h - d_h \in \mathbb{F}$, then FastTwoFMA gives:

$$e = \mathrm{RN}(ab + c_h - d_h) = \underbrace{ab + c_h}_{\in[1,2)} - d_h + \varepsilon_1$$

$$|ab + c_h - d_h| \leq u \Rightarrow |e| \leq u,\ |\varepsilon_1| \leq \frac{1}{2}u^2$$

FastTwoFMA_s : $\mathbb{F} \times \mathbb{F} + \mathrm{DW} \to \mathrm{DW}$



If $c_h - d_h \in \mathbb{F}$, then FastTwoFMA gives:

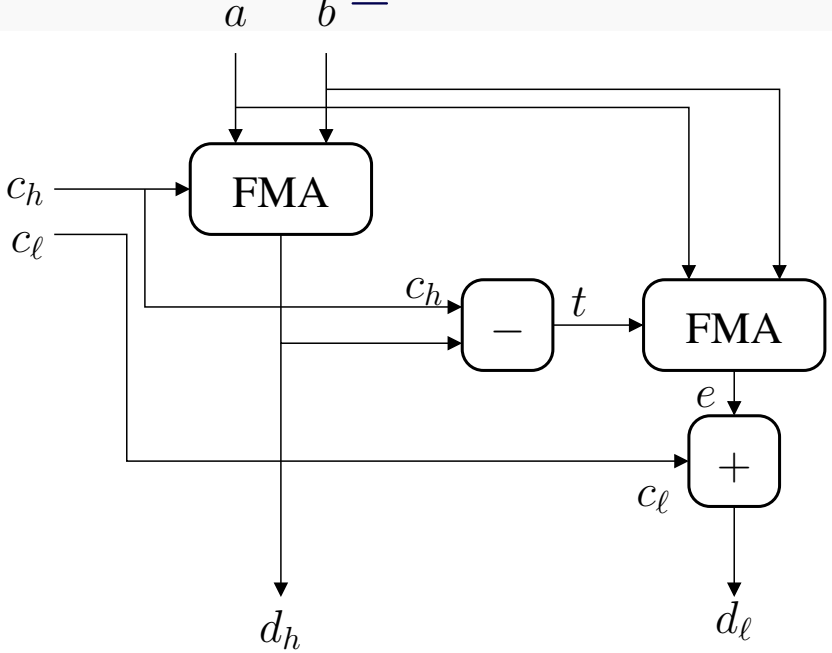
$$e = \mathrm{RN}(ab + c_h - d_h) = \underbrace{ab + c_h}_{\in[1,2)} - d_h + \varepsilon_1$$

$$|ab + c_h - d_h| \leq u \Rightarrow |e| \leq u,\ |\varepsilon_1| \leq \frac{1}{2}u^2$$

Low part:

$$d_\ell = \mathrm{RN}(e + c_\ell) = e + c_\ell + \varepsilon_2$$

FastTwoFMA_s : $\mathbb{F} \times \mathbb{F} + \mathrm{DW} \to \mathrm{DW}$



If $c_h - d_h \in \mathbb{F}$, then FastTwoFMA gives:

$$e = \mathrm{RN}(ab + c_h - d_h) = \underbrace{ab + c_h}_{\in[1,2)} - d_h + \varepsilon_1$$

$$|ab + c_h - d_h| \leq u \Rightarrow |e| \leq u,\ |\varepsilon_1| \leq \frac{1}{2}u^2$$

Parameterized overlap:

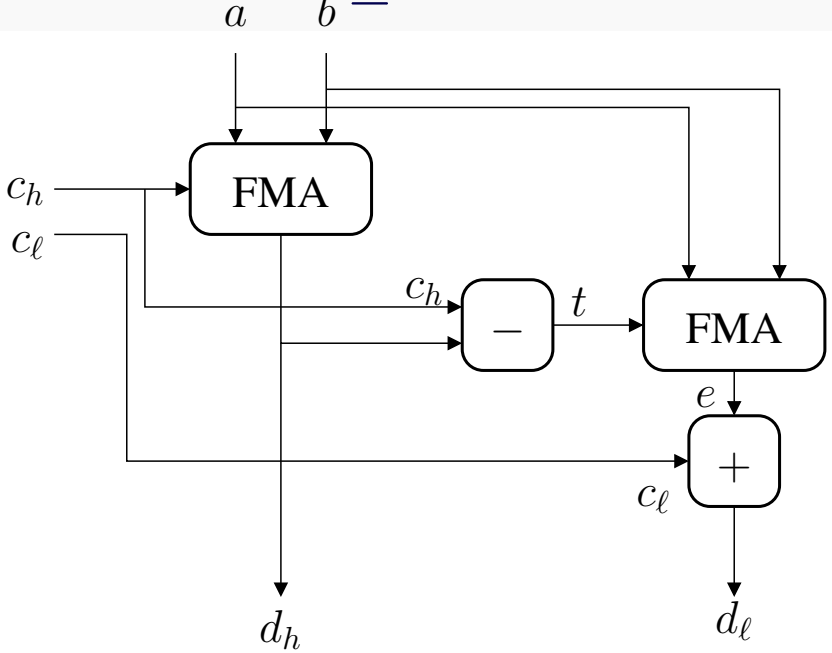
$$c = c_h + c_\ell, \quad |c_\ell| \leq k\,\mathrm{ulp}(c_h)$$

$$|c_h| \geq 2|ab| \Rightarrow |c_h| < 4, \quad |c_\ell| \leq 4ku$$

Low part:

$$d_\ell = \mathrm{RN}(e + c_\ell) = e + c_\ell + \varepsilon_2$$

FastTwoFMA_s : $\mathbb{F} \times \mathbb{F} + \mathrm{DW} \to \mathrm{DW}$



If $c_h - d_h \in \mathbb{F}$, then FastTwoFMA gives:

$$e = \mathrm{RN}(ab + c_h - d_h) = \underbrace{ab + c_h}_{\in[1,2)} - d_h + \varepsilon_1$$

$$|ab + c_h - d_h| \leq u \Rightarrow |e| \leq u,\ |\varepsilon_1| \leq \frac{1}{2}u^2$$

Parameterized overlap:

$$c = c_h + c_\ell, \quad |c_\ell| \leq k\,\mathrm{ulp}(c_h)$$

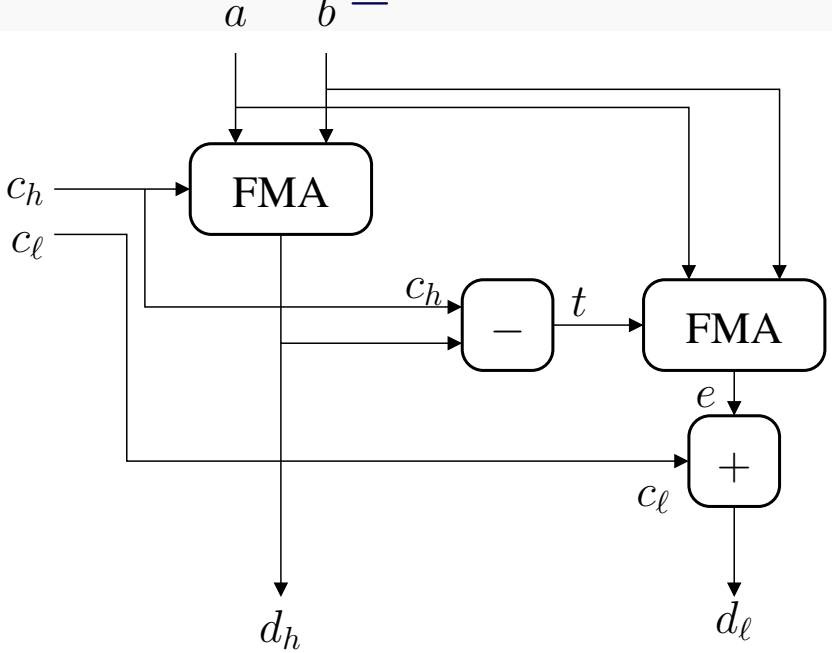
$$|c_h| \geq 2|ab| \Rightarrow |c_h| < 4, \quad |c_\ell| \leq 4ku$$

Low part:

$$d_\ell = \mathrm{RN}(e + c_\ell) = e + c_\ell + \varepsilon_2$$

$$|e + c_\ell| \leq (4k+1)u$$

FastTwoFMA_s : $\mathbb{F} \times \mathbb{F} + \mathrm{DW} \to \mathrm{DW}$



If $c_h - d_h \in \mathbb{F}$, then FastTwoFMA gives:

$$e = \mathrm{RN}(ab + c_h - d_h) = \underbrace{ab + c_h}_{\in[1,2)} - d_h + \varepsilon_1$$

$$|ab + c_h - d_h| \leq u \Rightarrow |e| \leq u,\ |\varepsilon_1| \leq \frac{1}{2}u^2$$

Parameterized overlap:

$$c = c_h + c_\ell, \quad |c_\ell| \leq k\,\mathrm{ulp}(c_h)$$

$$|c_h| \geq 2|ab| \Rightarrow |c_h| < 4, \quad |c_\ell| \leq 4ku$$

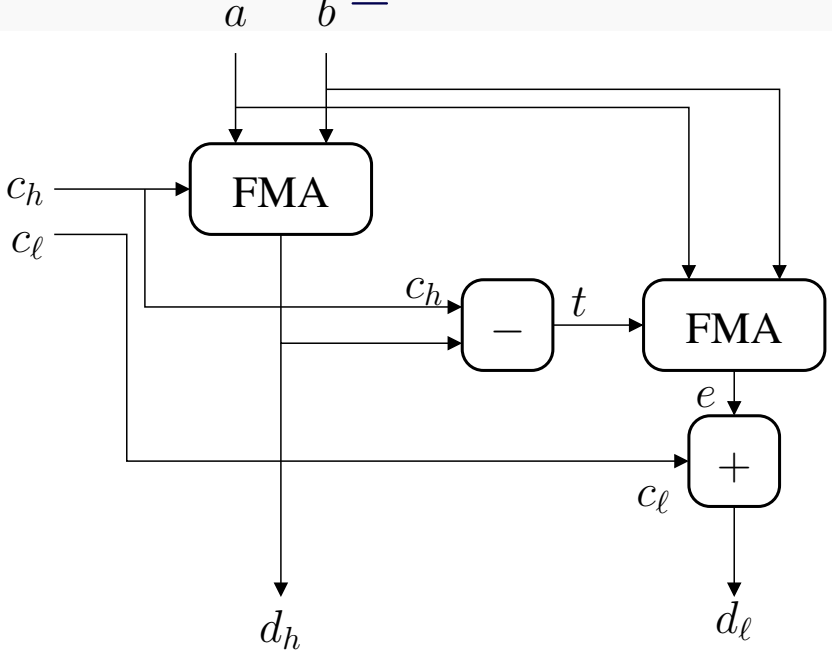
Low part:

$$d_\ell = \mathrm{RN}(e + c_\ell) = e + c_\ell + \varepsilon_2$$

$$|e + c_\ell| \leq (4k+1)u$$

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

FastTwoFMA_s : $\mathbb{F} \times \mathbb{F} + \mathrm{DW} \to \mathrm{DW}$



If $c_h - d_h \in \mathbb{F}$, then FastTwoFMA gives:

$$e = \mathrm{RN}(ab + c_h - d_h) = \underbrace{ab + c_h}_{\in[1,2)} - d_h + \varepsilon_1$$

$$|ab + c_h - d_h| \leq u \Rightarrow |e| \leq u,\ |\varepsilon_1| \leq \frac{1}{2}u^2$$

Parameterized overlap:

$$c = c_h + c_\ell, \quad |c_\ell| \leq k\,\mathrm{ulp}(c_h)$$

$$|c_h| \geq 2|ab| \Rightarrow |c_h| < 4, \quad |c_\ell| \leq 4ku$$

Low part:

$$d_\ell = \mathrm{RN}(e + c_\ell) = e + c_\ell + \varepsilon_2$$

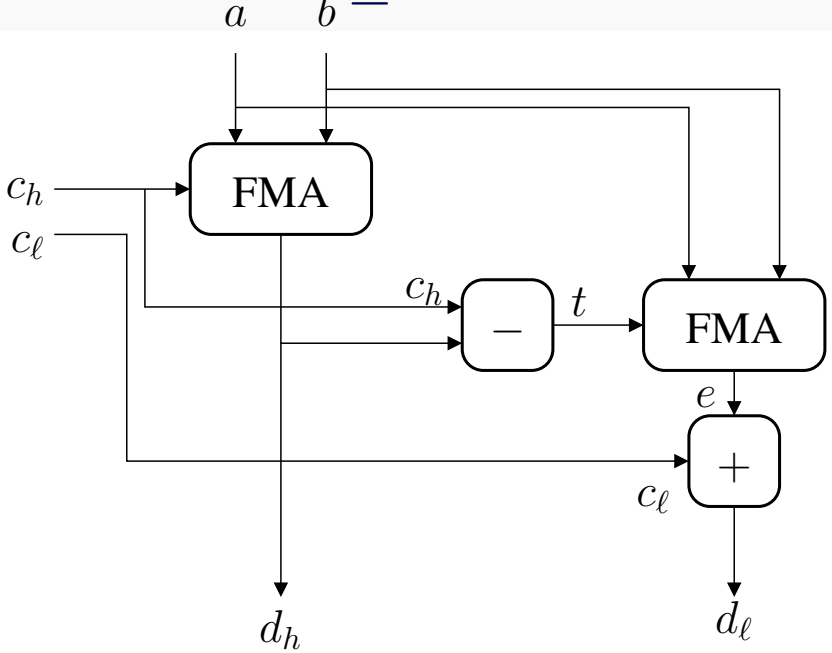
$$|e + c_\ell| \leq (4k+1)u$$

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

Error:

$$|d - (ab + c)| \leq |\varepsilon_1| + |\varepsilon_2|$$
$$\leq \ldots$$

FastTwoFMA_s : $\mathbb{F} \times \mathbb{F} + \mathrm{DW} \to \mathrm{DW}$



If $c_h - d_h \in \mathbb{F}$, then FastTwoFMA gives:

$$e = \mathrm{RN}(ab + c_h - d_h) = \underbrace{ab + c_h}_{\in[1,2)} - d_h + \varepsilon_1$$

$$|ab + c_h - d_h| \leq u \Rightarrow |e| \leq u,\ |\varepsilon_1| \leq \frac{1}{2}u^2$$

Parameterized overlap:

$$c = c_h + c_\ell, \quad |c_\ell| \leq k\,\mathrm{ulp}(c_h)$$

$$|c_h| \geq 2|ab| \Rightarrow |c_h| < 4, \quad |c_\ell| \leq 4ku$$

Low part:

$$d_\ell = \mathrm{RN}(e + c_\ell) = e + c_\ell + \varepsilon_2$$

$$|e + c_\ell| \leq (4k+1)u$$

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

Error:

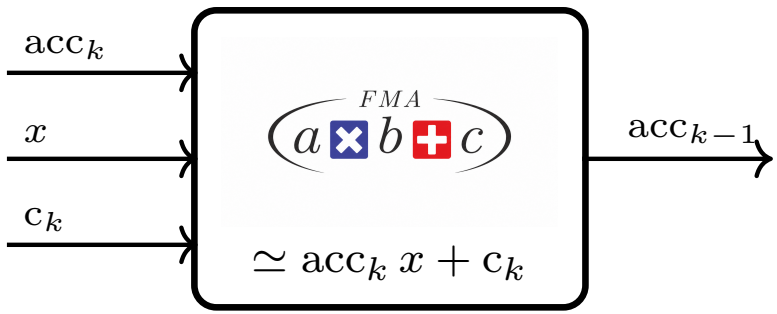
$$|d - (ab + c)| \leq |\varepsilon_1| + |\varepsilon_2|$$
$$\leq \dots$$
$$d = (ab + c)(1 + \mathrm{Const}_k\, u^2)$$

Motivation: a recurring kernel used in **libms**

Example: CORE-MATH accurate path for **binary64** exp

Degree-6 polynomial evaluation, Horner step:

$$\mathrm{acc}_{k-1} = \mathrm{acc}_k\, x + c_k, \qquad \mathrm{acc}_k, x, c_k \in \mathrm{DW}\,.$$



Classical DW-Horner step

DW multiplication followed by DW addition:

$$\mathrm{DDmul}(\mathrm{acc}_k, x) \quad + \quad \mathrm{DDadd}(\cdot, c_k).$$

$$3 \text{ mul} + 1 \text{ fma} + 7 \text{ add/sub}$$

per Horner step.

Improved DW-Horner step

Replacing with FastFMA_DW

$$\mathrm{acc}_{k-1} = \mathsf{FastFMA\_DW}(\mathrm{acc}_k, x, c_k).$$

$$4 \text{ fma} + 2 \text{ add/sub}$$

per Horner step.

Measured effect on hard-to-round cases

Benchmark subset

Hard-to-round cases that trigger the accurate path of CORE-MATH binary64 `exp`:

$106{,}471$ inputs from `exp.wc`, exponent in $[-4, 8]$.

CORE-MATH accurate path	latency	reciprocal throughput
Original `exp`	83.5	53.5
Modified with FastFMA_DW	73.7	47.4
Improvement	11.7%	11.4%

Takeaway

Horner step: classical/optimized DW mul–add $\rightsquigarrow$ FastFMA_DW
$\rightsquigarrow$ latency and throughput cost reduction on the hard-to-round accurate path.

Measurements on an Intel Core i5-1145G7, 2.60 GHz, using `gcc` 11.4.0/`glibc` 2.35 and the `RDTSC` counter; reported values are means over 20 trials, in cycles per call.

Remark: connection with vector correctly-rounded exp

The same FMA-based DW kernels are also useful in another ARITH 2026 paper

Correctly rounded vector implementation of the exponential function in binary64 arithmetic
Nicolas Brisebarre, Tom Hubrecht, Christoph Lauter, Jean-Michel Muller, Kristalys Ruiz-Rohena

Preliminary implementation feedback from Tom Hubrecht:

Remark: connection with vector correctly-rounded exp

The same FMA-based DW kernels are also useful in another ARITH 2026 paper

Correctly rounded vector implementation of the exponential function in binary64 arithmetic
Nicolas Brisebarre, Tom Hubrecht, Christoph Lauter, Jean-Michel Muller, Kristalys Ruiz-Rohena

Preliminary implementation feedback from Tom Hubrecht:

- Another mixed variant appears in practice – the addend is a single floating-point number:

$$\mathrm{DW} \times \mathrm{DW} + \mathbb{F} \longrightarrow \mathrm{DW}\,.$$

Remark: connection with vector correctly-rounded exp

The same FMA-based DW kernels are also useful in another ARITH 2026 paper

Correctly rounded vector implementation of the exponential function in binary64 arithmetic
Nicolas Brisebarre, Tom Hubrecht, Christoph Lauter, Jean-Michel Muller, Kristalys Ruiz-Rohena

Preliminary implementation feedback from Tom Hubrecht:

- Another mixed variant appears in practice – the addend is a single floating-point number:

$$\mathrm{DW} \times \mathrm{DW} + \mathbb{F} \longrightarrow \mathrm{DW}\,.$$

- Performance of a merged prototype for the vectorized exp:
  cycle count reduction by about

$$22\%$$

  on an AVX2 laptop.

  On AVX512 the gain is smaller, but the implementation remains competitive with highly optimized vector code.

Limitations and perspectives

1. The dominance condition is essential. If $|c_h| < 2|a_h b_h|$:
   - same-sign, non-cancelling cases can often be adapted;
   - near cancellation can make the fast kernel fail badly.

Example of failure under cancellation

For $a = (1, -u/4)$, $b = (1, u/2)$, $c = (-1, -u/4)$, FastFMA_DW returns 0 while the exact value is $-u^2/8$.

Limitations and perspectives

1. The dominance condition is essential. If $|c_h| < 2|a_h b_h|$:
   - same-sign, non-cancelling cases can often be adapted;
   - near cancellation can make the fast kernel fail badly.

Example of failure under cancellation

For $a = (1, -u/4)$, $b = (1, u/2)$, $c = (-1, -u/4)$, FastFMA_DW returns 0 while the exact value is $-u^2/8$.

2. Evaluate the impact of our kernels on other designs – especially on vector implementations where the fast-path often contains DW Horner evaluations.

Limitations and perspectives

1. The dominance condition is essential. If $|c_h| < 2|a_h b_h|$:
   - same-sign, non-cancelling cases can often be adapted;
   - near cancellation can make the fast kernel fail badly.

Example of failure under cancellation

For $a = (1, -u/4)$, $b = (1, u/2)$, $c = (-1, -u/4)$, FastFMA_DW returns 0 while the exact value is $-u^2/8$.

2. Evaluate the impact of our kernels on other designs – especially on vector implementations where the fast-path often contains DW Horner evaluations.

3. (Automatic) formal proofs of the bounds.

Takeaways

1. FMA-based extended-precision kernels are efficient under dominance.
2. DW overlap can be parameterized explicitly by k_a, k_b, k_c.
3. The paper gives tight or asymptotically tight constants for low-part size and relative error.
4. CORE-MATH `exp` case study: about 11% gain on tested hard cases.

Questions?

Backup: what improves Ozaki–Koizumi?

FastTwoFMA : $\mathbb{F} \times \mathbb{F} + \mathbb{F} \to \mathrm{DW}$

The algorithm is that of Ozaki–Koizumi:

$$\begin{aligned} d_h &\leftarrow \mathrm{RN}(ab + c), \\ t &\leftarrow \mathrm{RN}(c - d_h), \\ d_\ell &\leftarrow \mathrm{RN}(ab + t). \end{aligned}$$

Their analysis gives, under $c - d_h \in \mathbb{F}$,

$$|\delta| \leq \frac{u^2}{(1+u)^2} < u^2, \qquad |d_\ell| \leq u\,|ab + c|.$$

Our result sharpens this to

$$\boxed{|\delta| < \frac{1}{2}u^2}, \qquad \boxed{|d_\ell| \leq \frac{1}{2}\,\mathrm{ulp}(d_h)}.$$

Moreover:

$\frac{1}{2}u^2$ is asymptotically optimal, $\frac{1}{2}\,\mathrm{ulp}(d_h)$ is optimal.

Backup: what improves Ozaki–Koizumi for FastTwoFMA_s?

FastTwoFMA_s : $\mathbb{F} \times \mathbb{F} + \mathrm{DW} \to \mathrm{DW}$

The algorithm is again from Ozaki–Koizumi:

$$\begin{aligned} d_h &\leftarrow \mathrm{RN}(ab + c_h), \\ t &\leftarrow \mathrm{RN}(c_h - d_h), \\ e &\leftarrow \mathrm{RN}(ab + t), \\ d_\ell &\leftarrow \mathrm{RN}(e + c_\ell). \end{aligned}$$

Assume

$$|c_h| \geq 2|ab|, \qquad |c_\ell| \leq k_c \ \mathrm{ulp}(c_h).$$

Previous error analysis

Ozaki–Koizumi give rigorous bounds essentially of the form

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

and

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

A direct refinement using the improved FastTwoFMA theorem gives

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

For classical DW overlap this implies only

$$|\delta| \lesssim 6.5u^2.$$

Backup: general FastFMA_DW relative-error theorem

Let

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

If $|c_h| \ge 2|a_h b_h|$, then

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

where

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

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