Rounding Errors Statistics as Numerics Signature

Ping Tak Peter Tang
Rivos Inc.
May 6, 2025
ARITH 2025, El Paso, Texas

Rounding Errors Stats as Numerics Signature

- Motivation
- Proposition
- Validation
- Application

Rounding Errors Stats as Numerics Signature

- Motivation
- Proposition
- Validation
- Application

Motivation

- Proliferation of non-standard arithmetic
- Threshold test on single error ineffective
- Need more robust alternatives

Threshold Test/Single Error – An Example

```
def sgl_simd_sum(x_data, simd_len):
    L = len(x_data); s = 0.0; ind = 0;
    n_outer = L // simd_len
    for _ in range(n_outer):
        tmp = 0.0
        for _ in range(simd_len):
            tmp = tmp + x_data[ind]
            ind += 1
        s = s + tmp
    return s
```

Threshold Test/Single Error – An Example

```
def sgl_simd_sum(x_data, simd_len):
    L = len(x_data); s = 0.0; ind = 0;
    n_outer = L // simd_len
    for _ in range(n_outer):
        tmp = 0.0
        for _ in range(simd_len):
            tmp = tmp + x_data[ind]
            ind += 1
        s = s + tmp
    return s
```

Threshold Test/Single Error – An Example

```
def sgl_simd_sum(x_data, simd_len):
    L = len(x_data); s = 0.0; ind = 0;
    n_outer = L // simd_len
    for _ in range(n_outer):
        tmp = 0.0
        for _ in range(simd_len):
            tmp = tmp + x_data[ind]
            ind += 1
        s = s + tmp
    return s
```

Threshold Test/Single Error – An Example

Expect kernel is `simd_len == 4`
Generate $L = 64$ random inputs
compute error E (vs. double)
Expect $|E| < 18\epsilon$

```
def sgl_simd_sum(x_data, simd_len):
    L = len(x_data); s = 0.0; ind = 0;
    n_outer = L // simd_len
    for _ in range(n_outer):
        tmp = 0.0
        for _ in range(simd_len):
            tmp = tmp + x_data[ind]
            ind += 1
        s = s + tmp
    return s
```

Threshold Test/Single Error – An Example

Expect kernel is `simd_len == 4`
Generate $L = 64$ random inputs
compute error E (vs. double)
Expect $|E| < 18\epsilon$

Fails because $E = -18.7\epsilon$
Change theshold to 19ϵ
ALL IS WELL

```
def sgl_simd_sum(x_data, simd_len):
    L = len(x_data); s = 0.0; ind = 0;
    n_outer = L // simd_len
    for _ in range(n_outer):
        tmp = 0.0
        for _ in range(simd_len):
            tmp = tmp + x_data[ind]
            ind += 1
        s = s + tmp
    return s
```

Threshold Test/Single Error – An Example

Expect kernel is `simd_len == 4`
Generate $L = 64$ random inputs
compute error E (vs. double)
Expect $|E| < 18\epsilon$

Fails because $E = -18.7\epsilon$
Change theshold to 19ϵ
ALL IS WELL

until

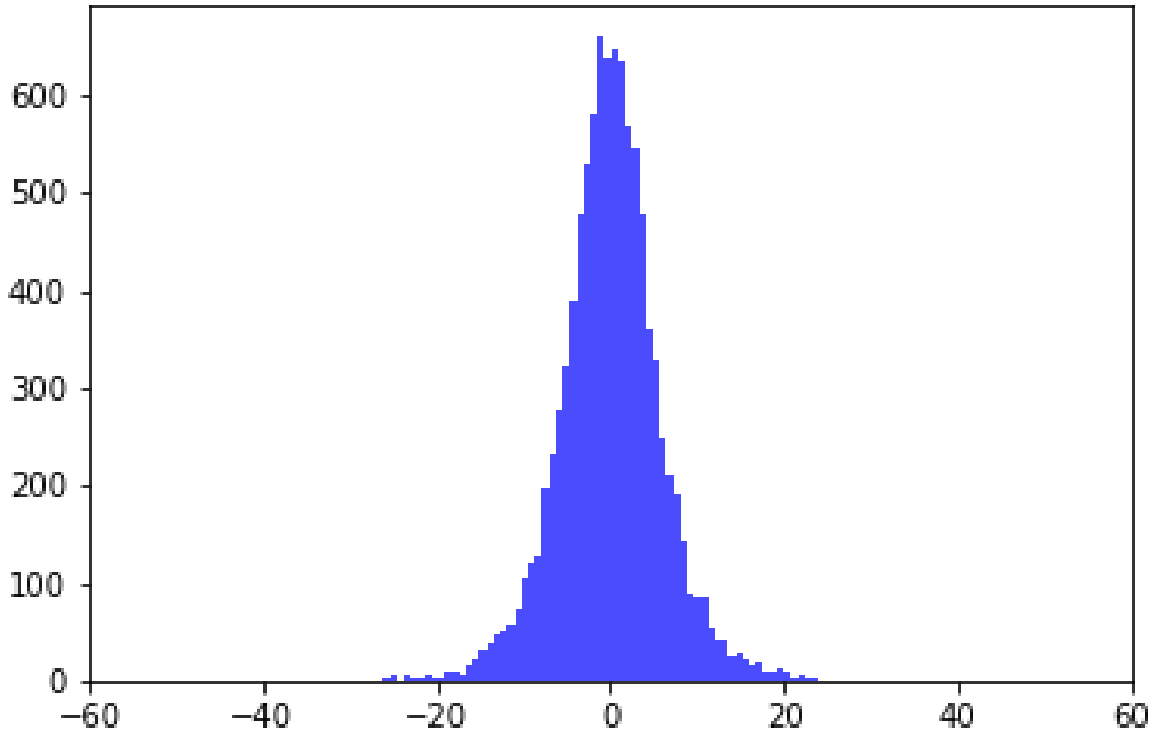
```
def sgl_simd_sum(x_data, simd_len):
    L = len(x_data); s = 0.0; ind = 0;
    n_outer = L // simd_len
    for _ in range(n_outer):
        tmp = 0.0
        for _ in range(simd_len):
            tmp = tmp + x_data[ind]
            ind += 1
        s = s + tmp
    return s
```

Threshold Test/Single Error – An Example

Expect kernel is `simd_len == 4`
Generate $L = 64$ random inputs
compute error E (vs. double)
Expect $|E| < 18\epsilon$

Fails because $E = -18.7\epsilon$
Change theshold to 19ϵ
ALL IS WELL

until



histogram of 10^6 error (units ϵ) simd 4

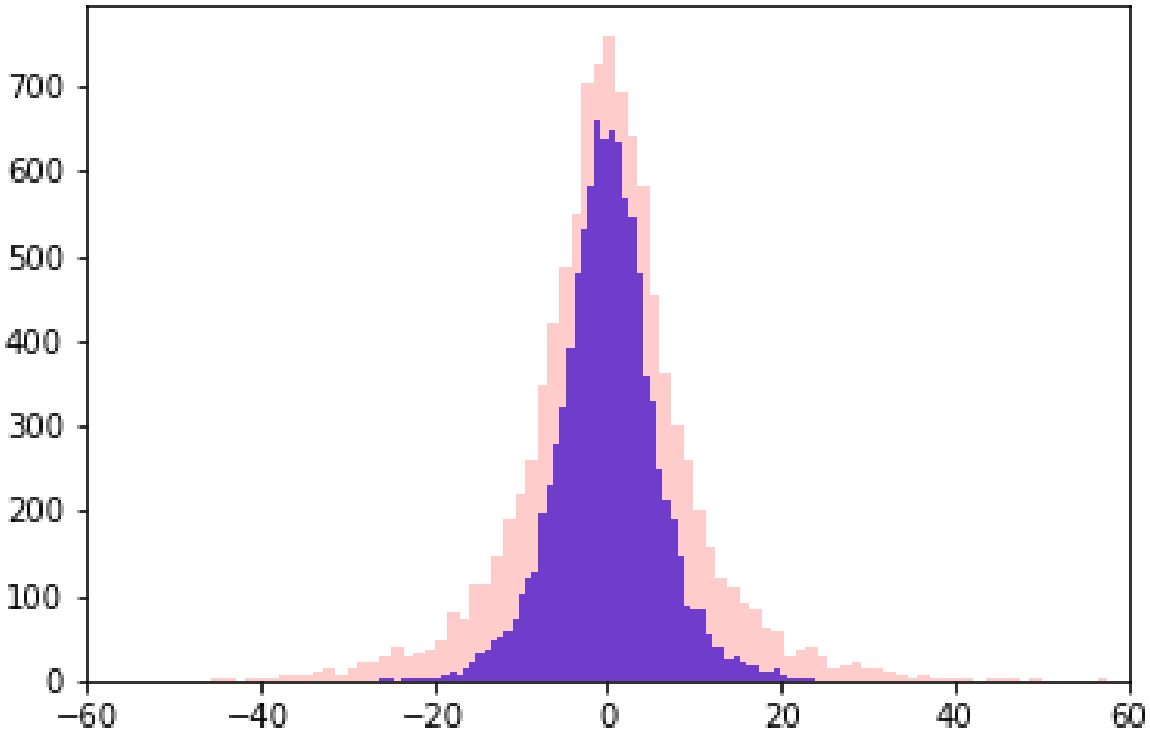
Threshold Test/Single Error – An Example

Expect kernel is `simd_len == 4`
Generate $L = 64$ random inputs
compute error E (vs. double)
Expect $|E| < 18\epsilon$

Fails because $E = -18.7\epsilon$
Change theshold to 19ϵ
ALL IS WELL

until

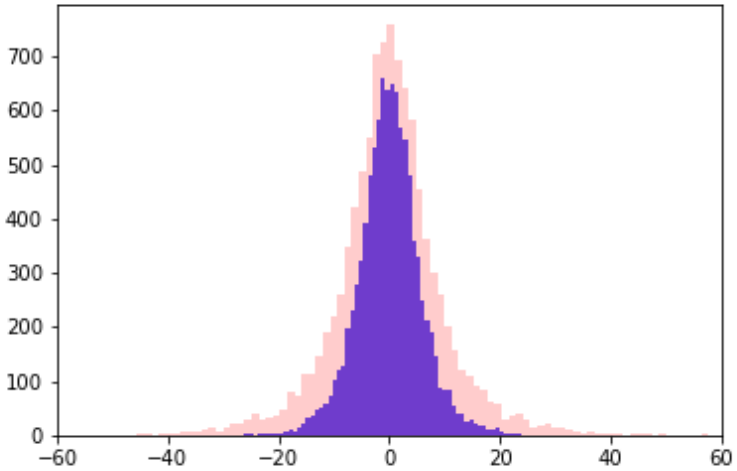
histogram of 10^6 error (units ϵ) simd 1



histogram of 10^6 error (units ϵ) simd 4

Proposition

- Use statistics instead of a single error to judge numerics
- Belief: computational errors has specific distributions, and
- Statistics such as variance has a narrow spread



Validation

1. Use specific input distributions
2. Model distribution of computational errors
3. Match sampled variance to theoretical values for several computational kernels

Validation

1. Use specific input distributions
2. Model distribution of computational errors
3. Match sampled variance to theoretical values for several computational kernels
    1. SIMD summation
    2. Inner products
    3. Matrix multiplication
    4. Hybrid fixed-float summation

Validation

1. Use specific input distributions
2. Model distribution of computational errors
3. Match sampled variance to theoretical values for several computational kernels

1. SIMD summation		***here***
2. Inner products 3. Matrix multiplication	***see paper***	
4. Hybrid fixed-float summation		***here***

Validation

Model Basic Rounding

Take $x \in \mathbb{R}$, $x \sim \mathcal{N}(0, \sigma^2)$. Define $\mathcal{R}_\sigma := fl(x) - x$.

Validation

Model Basic Rounding

Take $x \in \mathbb{R}$, $x \sim \mathcal{N}(0, \sigma^2)$. Define $\mathcal{R}_\sigma := fl(x) - x$.

If $|x| \in I_k = [2^k, 2^{k+1})$, then $\mathcal{R}_\sigma \sim 2^k \mathcal{U}[-\frac{\epsilon}{2}, \frac{\epsilon}{2}]$, $\epsilon = ulp(1)$.

Validation

Model Basic Rounding

Take $x \in \mathbb{R}$, $x \sim \mathcal{N}(0, \sigma^2)$. Define $\mathcal{R}_\sigma := fl(x) - x$.

If $|x| \in I_k = [2^k, 2^{k+1})$, then $\mathcal{R}_\sigma \sim 2^k \mathcal{U}[-\frac{\epsilon}{2}, \frac{\epsilon}{2}]$, $\epsilon = ulp(1)$.

Because $\mathrm{Var}(\mathcal{U}[-\frac{\epsilon}{2}, \frac{\epsilon}{2}]) = \frac{\epsilon^2}{12}$

and probability of $|x| \in I_k$ is $\quad p_k(\sigma) = \frac{2}{\sqrt{2\pi\sigma^2}} \int_{2^k}^{2^{k+1}} e^{-\frac{t^2}{2\sigma^2}} dt$

Thus $E(\mathcal{R}_\sigma^2) = \frac{\epsilon^2}{12} \sum 2^{2k} p_k(\sigma) = \frac{\epsilon^2}{12} F(\sigma) = \frac{\epsilon^2}{12} \sigma^2 F_0(\sigma)$
(easy to see $F_0(2\sigma) = F_0(\sigma)$)

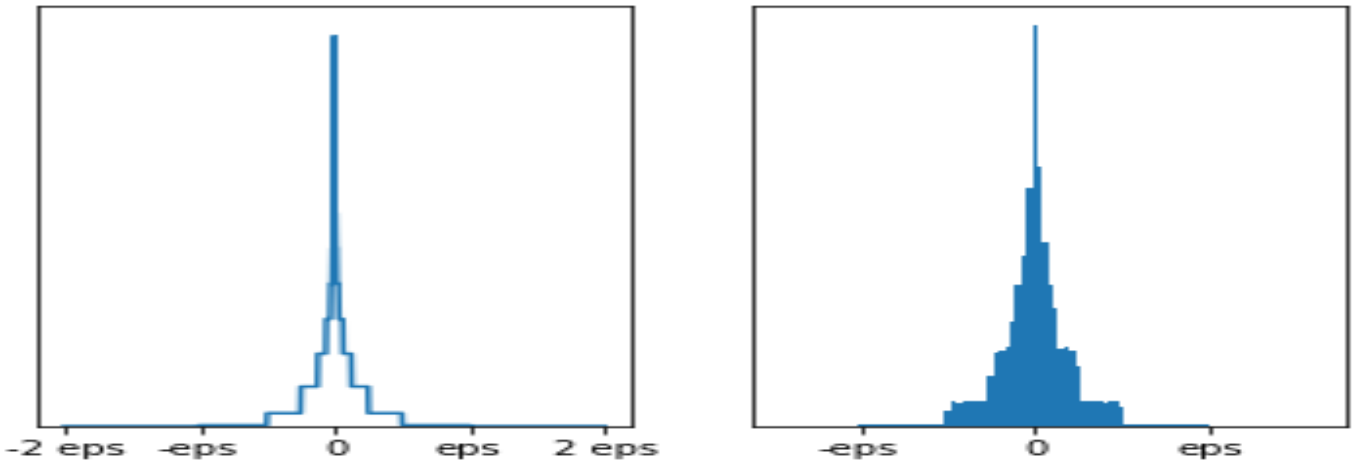
Validation

Model Basic Rounding

Take $x \in \mathbb{R}$, $x \sim \mathcal{N}(0, \sigma^2)$. Define $\mathcal{R}_\sigma := fl(x) - x$.

If $|x| \in I_k = [2^k, 2^{k+1})$, then $\mathcal{R}_\sigma \sim 2^k \mathcal{U}[-\frac{\epsilon}{2}, \frac{\epsilon}{2}]$, $\epsilon = ulp(1)$.

Probability density function and histogram of
$\mathcal{R}_\sigma$, $x \sim \mathcal{N}(0, 1)$, $\epsilon = 2^{-23}$



Validation

Model Basic Rounding

Take $x \in \mathbb{R}$, $x \sim \mathcal{N}(0, \sigma^2)$. Define $\mathcal{R}_\sigma := fl(x) - x$.

If $|x| \in I_k = [2^k, 2^{k+1})$, then $\mathcal{R}_\sigma \sim 2^k \mathcal{U}[-\frac{\epsilon}{2}, \frac{\epsilon}{2}]$, $\epsilon = ulp(1)$.

Thus $E(\mathcal{R}_\sigma^2) = \frac{\epsilon^2}{12} \sum 2^{2k} p_k(\sigma) = \frac{\epsilon^2}{12} F(\sigma) = \frac{\epsilon^2}{12} \sigma^2 F_0(\sigma)$
(easy to see $F_0(2\sigma) = F_0(\sigma)$)

Similary $E(\mathcal{R}_\sigma^4) = \frac{\epsilon^4}{80} \sum 2^{4k} p_k(\sigma) = \frac{\epsilon^4}{80} G(\sigma) = \frac{\epsilon^4}{80} \sigma^4 G_0(\sigma)$
(easy to see $G_0(2\sigma) = G_0(\sigma)$)

Validation

Model Basic Rounding

Take $x \in \mathbb{R}$, $x \sim \mathcal{N}(0, \sigma^2)$. Define $\mathcal{R}_\sigma := fl(x) - x$.

$$E(\mathcal{R}_\sigma^2) = \frac{\epsilon^2}{12} \sum 2^{2k} p_k(\sigma) = \frac{\epsilon^2}{12} F(\sigma) = \frac{\epsilon^2}{12} \sigma^2 F_0(\sigma), \qquad F_0(2\sigma) = F_0(\sigma)$$

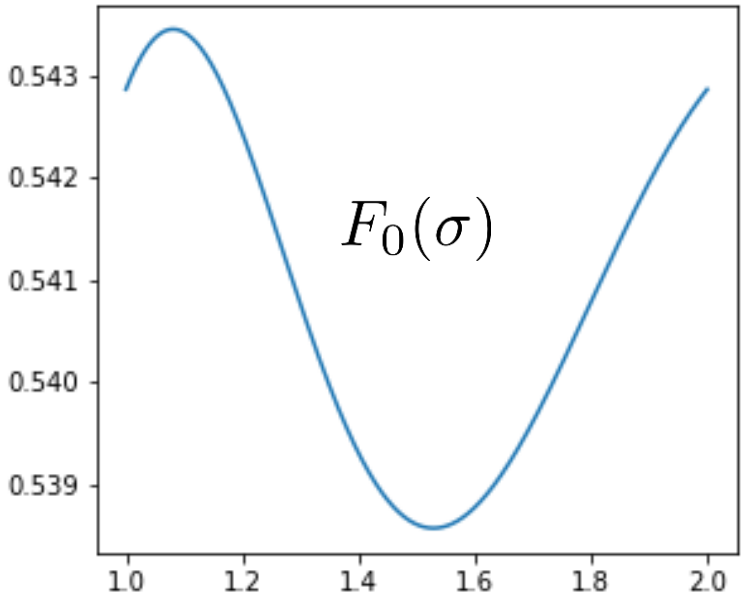
$$E(\mathcal{R}_\sigma^4) = \frac{\epsilon^4}{80} \sum 2^{4k} p_k(\sigma) = \frac{\epsilon^4}{80} G(\sigma) = \frac{\epsilon^4}{80} \sigma^4 G_0(\sigma), \qquad G_0(2\sigma) = G_0(\sigma)$$

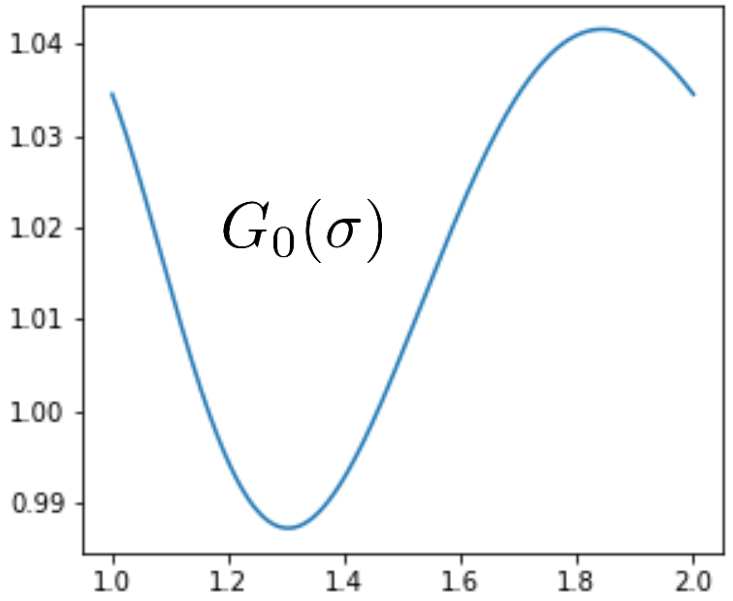
Validation

Model Basic Rounding

Take $x \in \mathbb{R}$, $x \sim \mathcal{N}(0, \sigma^2)$. Define $\mathcal{R}_\sigma := fl(x) - x$.

$$E(\mathcal{R}_\sigma^2) = \frac{\epsilon^2}{12}\sigma^2 F_0(\sigma) \qquad E(\mathcal{R}_\sigma^4) = \frac{\epsilon^4}{80}\sigma^4 G_0(\sigma)$$





Validation

Model Addition

$$x \sim \quad \mathcal{N}(0, \sigma_x^2) \quad , y \sim \quad \mathcal{N}(0, \sigma_y^2) \quad . \quad \sigma^2 = \sigma_x^2 + \sigma_y^2.$$

$$S_\sigma \quad = fl(x+y) - (x+y)$$

$$E(S_\sigma^2 \quad ||x| \in I_k) = \frac{\epsilon^2}{12} 2^{2k}$$

Validation

Model Addition

$$x \sim fl(\ \mathcal{N}(0, \sigma_x^2)\ )\ , y \sim fl(\ \mathcal{N}(0, \sigma_y^2)\ )\ . \quad \sigma^2 = \sigma_x^2 + \sigma_y^2.$$

$$S_\sigma \quad = fl(x+y) - (x+y)$$

$$E(S_\sigma^2 \ , \ |x| \in I_k) = \frac{\epsilon^2}{12} 2^{2k}$$

Validation

Model Addition

$x \sim fl(\ \mathcal{N}(0, \sigma_x^2)\ )\ , y \sim fl(\ \mathcal{N}(0, \sigma_y^2)\ )\ .\quad \sigma^2 = \sigma_x^2 + \sigma_y^2. \qquad r = \sigma_x^2/\sigma_y^2.$

$S_{\sigma\ ,r} = fl(x+y) - (x+y)$

$E(S_{\sigma\ ,r}^2 || x| \in I_k) = \frac{\epsilon^2}{12} 2^{2k}\ \ (\sum \alpha_{\pm,\ell} P((x/y) \in \pm I_\ell))$

x/y is a Cauchy distribution, only dependent on r, not σ

Validation

Model Addition

$x \sim fl(\ \mathcal{N}(0, \sigma_x^2)\ )\ , y \sim fl(\ \mathcal{N}(0, \sigma_y^2)\ )\ .\quad \sigma^2 = \sigma_x^2 + \sigma_y^2. \qquad r = \sigma_x^2/\sigma_y^2.$

$S_{\sigma\ ,r} = fl(x+y) - (x+y)$

$E(S_{\sigma\ ,r}^2 || x| \in I_k) = \frac{\epsilon^2}{12} 2^{2k}\ \ (\sum \alpha_{\pm,\ell} P((x/y) \in \pm I_\ell))$

x/y is a Cauchy distribution, only dependent on r, not σ

$E(\mathcal{S}_{\sigma,r}^2) = \frac{\epsilon^2}{12}\sigma^2 F_0(\sigma)\phi(r) \qquad E(\mathcal{S}_{\sigma,r}^4) = \frac{\epsilon^4}{80}\sigma^4 G_0(\sigma)\psi(r)$

Validation

Model Addition

$$x \sim fl(\,\mathcal{N}(0,\sigma_x^2)\,)\,, y \sim fl(\,\mathcal{N}(0,\sigma_y^2)\,)\,. \quad \sigma^2 = \sigma_x^2 + \sigma_y^2. \qquad r = \sigma_x^2/\sigma_y^2.$$

$$S_{\sigma,r} = fl(x+y) - (x+y) \qquad E(\mathcal{S}_{\sigma,r}^2) = \frac{\epsilon^2}{12}\sigma^2 F_0(\sigma)\phi(r) \qquad E(\mathcal{S}_{\sigma,r}^4) = \frac{\epsilon^4}{80}\sigma^4 G_0(\sigma)\psi(r)$$

	$F(\sigma)/\sigma^2$	$G(\sigma)/\sigma^4$	$\phi(r)$	$\psi(r)$
c_0	0.54279	1.03445	1.00684	0.99967
c_1	0.01827	-0.16159	0.88421	1.94703
c_2	-0.13577	-0.85648	-1.91853	-2.43573
c_3	0.16304	4.03253	-1.93557	5.04192
c_4	0.10800	-2.64662	-0.72735	-16.59014
c_5	-0.26253	-7.00256	0.00000	29.25313
c_6	0.10911	13.25400	0.00000	-24.83812
c_7	0.00000	-8.72614	0.00000	9.10214
c_8	0.00000	2.10696	0.00000	-0.84897

$F_0(\sigma), G_0(\sigma)$ are of the form $\sum_{j=0}^{d} c_j(\sigma-1)^j$, $1 \le \sigma \le 2$

$\phi(r), \psi(r)$ are of the form $\sum_{j=0}^{d} c_j r^j$, $0 \le r \le 1$.

Validation

Model SIMD Sum

$x_i \sim fl(\mathcal{N}(0,1))$, $\texttt{simd\_len} = \ell$, $m = L/\ell$

```
def sgl_simd_sum(x_data, simd_len):
    L = len(x_data); s = 0.0; ind = 0;
    n_outer = L // simd_len
    for _ in range(n_outer):
        tmp = 0.0
        for _ in range(simd_len):
            tmp = tmp + x_data[ind]
            ind += 1
        s = s + tmp
    return s
```

Validation

Model SIMD Sum

$x_i \sim fl(\mathcal{N}(0,1))$, $\texttt{simd\_len} = \ell$, $m = L/\ell$

m inner-loop sum of ℓ values, of the form
$a_0 = y_0$; $a_i = fl(a_{i-1} + y_i)$, $i = 1, 2, \ldots, \ell - 1$

Each error is $\mathcal{S}_{\sigma_i, r_i}$, $\sigma_i = \sqrt{1+i}$, $r_i = 1/i$

```
def sgl_simd_sum(x_data, simd_len):
    L = len(x_data); s = 0.0; ind = 0;
    n_outer = L // simd_len
    for _ in range(n_outer):
        tmp = 0.0
        for _ in range(simd_len):
            tmp = tmp + x_data[ind]
            ind += 1
        s = s + tmp
    return s
```

Validation

Model SIMD Sum

$x_i \sim fl(\mathcal{N}(0,1))$, $\texttt{simd\_len} = \ell$, $m = L/\ell$

m inner-loop sum of ℓ values, of the form
$a_0 = y_0$; $a_i = fl(a_{i-1} + y_i)$, $i = 1, 2, \ldots, \ell - 1$

Each error is $\mathcal{S}_{\sigma_i, r_i}$, $\sigma_i = \sqrt{1+i}$, $r_i = 1/i$

1 outer-loop sum of m values, of the form
$a_0 = y_0$; $a_i = fl(a_{i-1} + y_i)$, $i = 1, 2, \ldots, m - 1$

Each error is $\mathcal{S}_{\sigma_i, r_i}$, $\sigma_i = \sqrt{(1+i)\ell}$, $r_i = 1/i$

```
def sgl_simd_sum(x_data, simd_len):
    L = len(x_data); s = 0.0; ind = 0;
    n_outer = L // simd_len
    for _ in range(n_outer):
        tmp = 0.0
        for _ in range(simd_len):
            tmp = tmp + x_data[ind]
            ind += 1
        s = s + tmp
    return s
```

Validation

Model SIMD Sum

$x_i \sim fl(\mathcal{N}(0,1))$, $\texttt{simd\_len} = \ell$, $m = L/\ell$

Inner-loop error $\mathcal{S}_{\sigma_i, r_i}$, $\sigma_i = \sqrt{1+i}$, $r_i = 1/i$

$i = 1, 2, \ldots, \ell - 1$, m sets

Outer-loop error $\mathcal{S}_{\sigma_i, r_i}$, $\sigma_i = \sqrt{(1+i)\ell}$, $r_i = 1/i$

$i = 1, 2, \ldots, \ell - 1$, one set

Total error $\texttt{SIMD} - \ell$ sum: Δ_ℓ

$\Delta_{(\ell)} = \tau_1 + \tau_2 + \cdots + \tau_{L-1}$

τ_i, τ_j independent and $E(\tau_i) = 0$

$$E(\Delta_\ell^2) = \frac{\epsilon^2}{12}\left(m\sum_{i=1}^{\ell-1} F(\sqrt{1+i})\phi(1/i) + \sum_{i=1}^{m-1} F(\sqrt{1+i})\phi(1/i)\right)$$

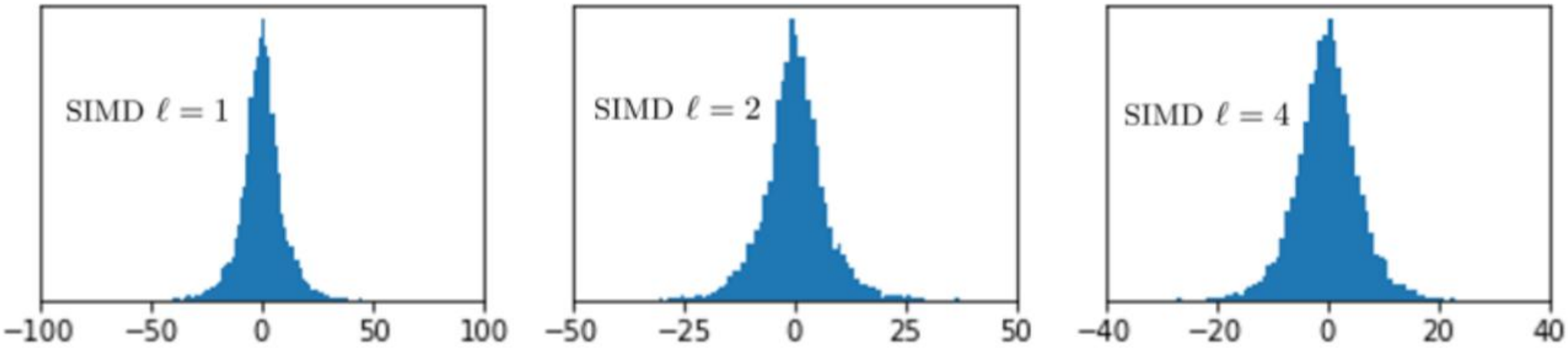
```
def sgl_simd_sum(x_data, simd_len):
    L = len(x_data); s = 0.0; ind = 0;
    n_outer = L // simd_len
    for _ in range(n_outer):
        tmp = 0.0
        for _ in range(simd_len):
            tmp = tmp + x_data[ind]
            ind += 1
        s = s + tmp
    return s
```

Validation

Model SIMD Sum

$N = 10^4$ summation experiments

Histogram of $\Delta_{(\ell),i}/\epsilon$, summation with SIMD addition. $\epsilon = 2^{-23}$



Variance: $96.1\epsilon^2$ (theory $96.7\epsilon^2$)

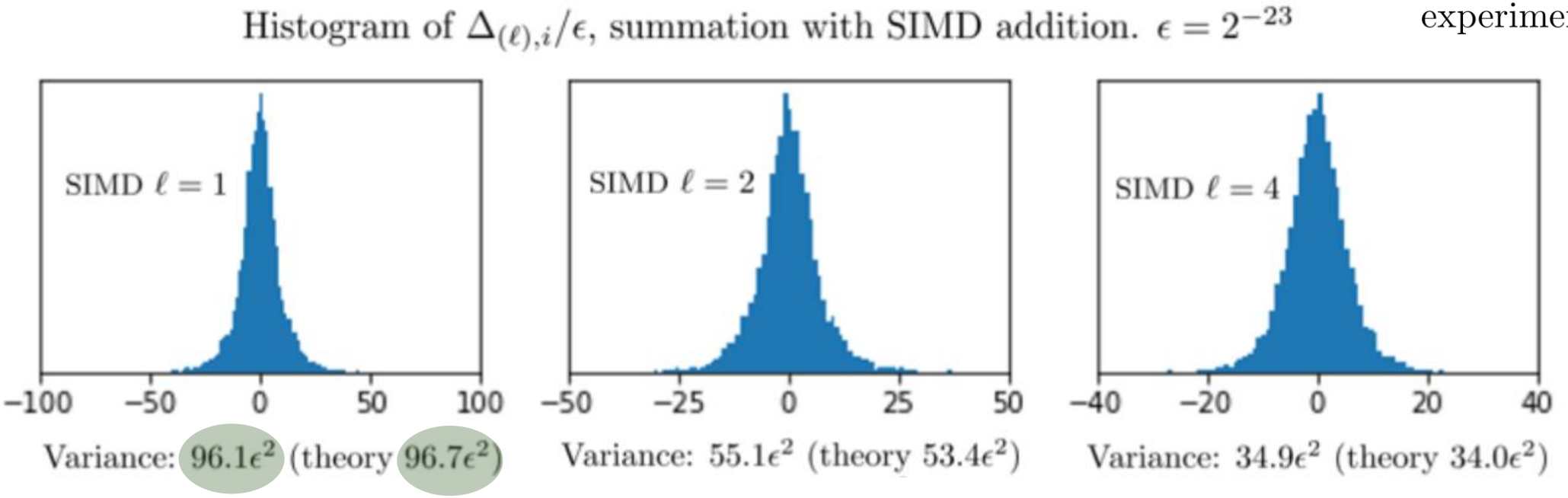
Variance: $55.1\epsilon^2$ (theory $53.4\epsilon^2$)

Variance: $34.9\epsilon^2$ (theory $34.0\epsilon^2$)

Validation

Model SIMD Sum

$N = 10^4$ summation experiments

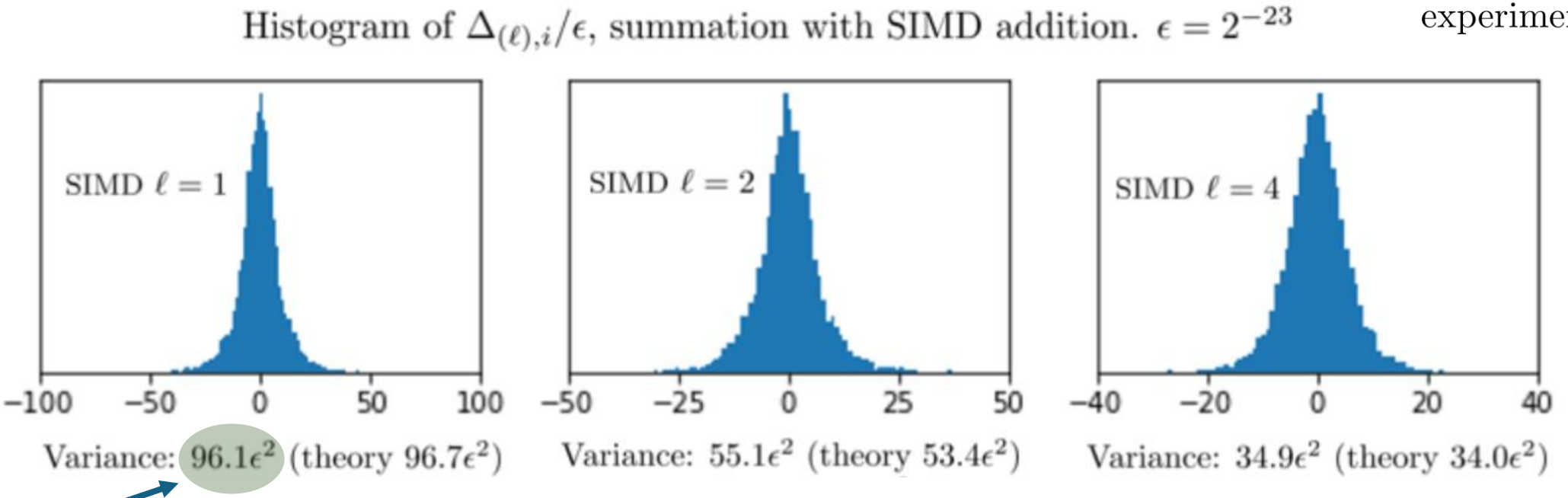


is the difference reasonable?

Validation

Model SIMD Sum

$N = 10^4$ summation experiments



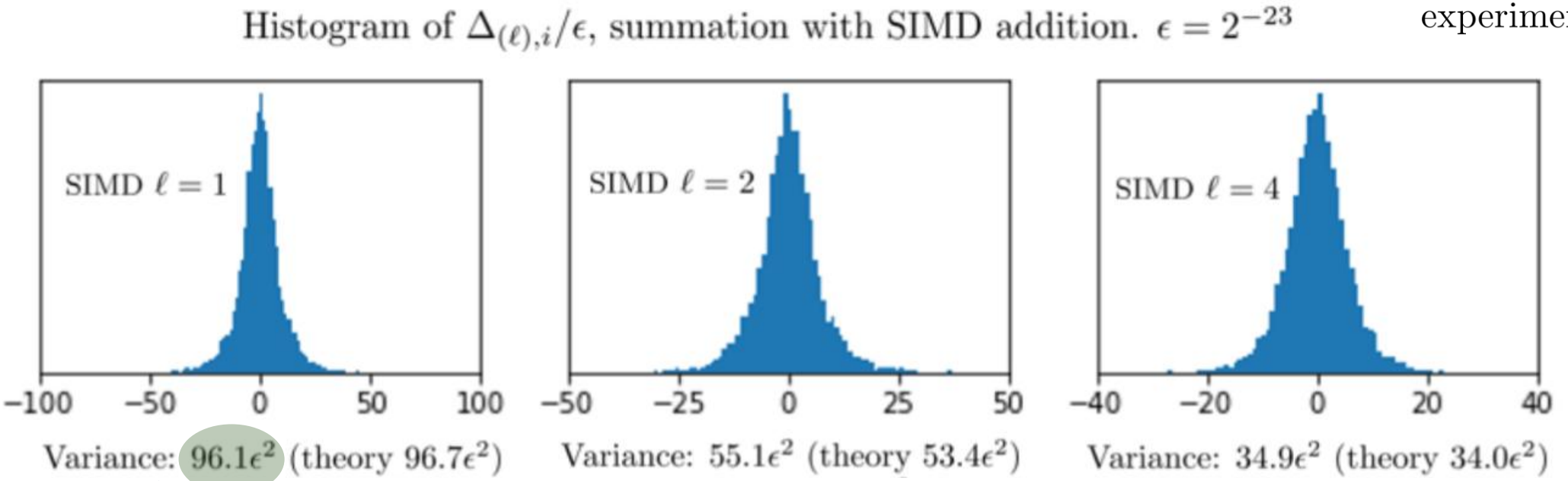
One sample of $Z = \sum_{i=1}^{N} \Delta_{(\ell),i}^2/N$

$\Delta_{(\ell)} = \tau_1 + \tau_2 + \cdots + \tau_{L-1}$

Validation

Model SIMD Sum

$N = 10^4$ summation experiments



One sample of $Z = \sum_{i=1}^{N} \Delta_{(\ell),i}^2/N \sim \mathrm{Var}(\Delta_{(\ell)}) + \frac{\sigma}{\sqrt{N}}\mathcal{N}(0,1)$ $\quad \sigma^2 = \mathrm{Var}(\Delta_{(\ell)}^2)$

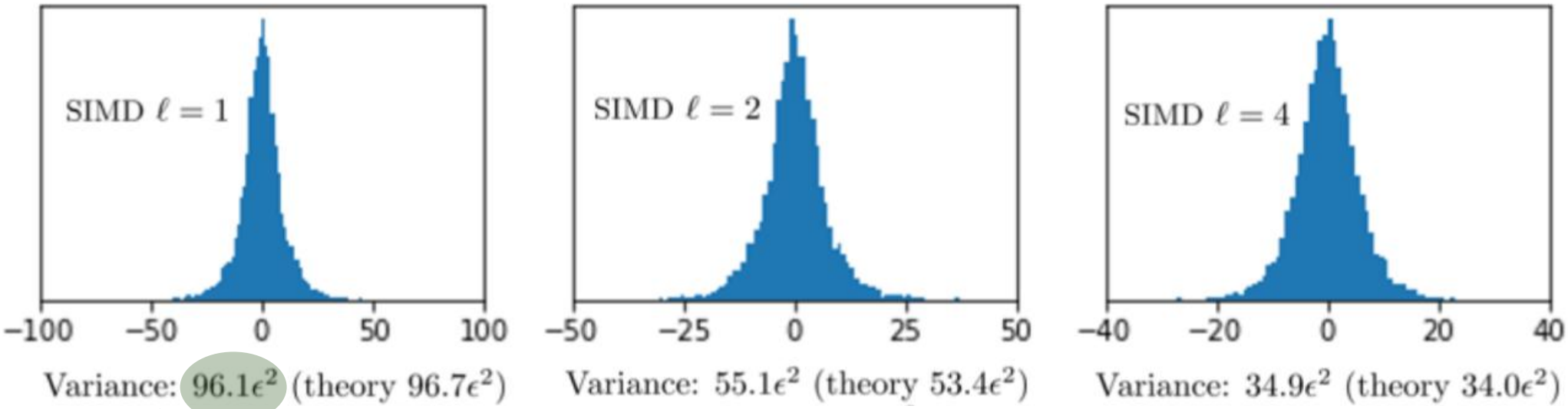
$\Delta_{(\ell)} = \tau_1 + \tau_2 + \cdots + \tau_{L-1}$

Validation

Model SIMD Sum

$N = 10^4$ summation experiments

Histogram of $\Delta_{(\ell),i}/\epsilon$, summation with SIMD addition. $\epsilon = 2^{-23}$



Variance: $96.1\epsilon^2$ (theory $96.7\epsilon^2$)

Variance: $55.1\epsilon^2$ (theory $53.4\epsilon^2$)

Variance: $34.9\epsilon^2$ (theory $34.0\epsilon^2$)

One sample of $Z = \sum_{i=1}^{N} \Delta_{(\ell),i}^2/N \sim \mathrm{Var}(\Delta_{(\ell)}) + \frac{\sigma}{\sqrt{N}}\mathcal{N}(0,1)$ $\quad \sigma^2 = \mathrm{Var}(\Delta_{(\ell)}^2)$

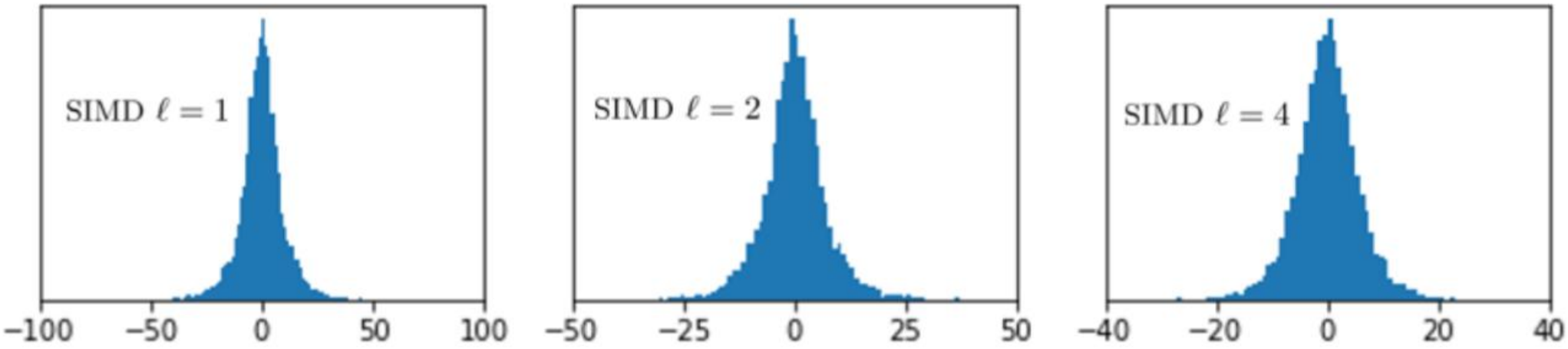
$\Delta_{(\ell)} = \tau_1 + \tau_2 + \cdots + \tau_{L-1}$

$\mathrm{Var}(\Delta_{(\ell)}^2) = E(\Delta_{(\ell)}^4) - E^2(\Delta_{(\ell)}^2) = \sum_i ((E(\tau_i^4) - E^2(\tau_i^2)) + 10 \sum_{i<j} E(\tau_i^2)E(\tau_j^2)$

Validation

Model SIMD Sum

Histogram of $\Delta_{(\ell),i}/\epsilon$, summation with SIMD addition. $\epsilon = 2^{-23}$



Variance: $96.1\epsilon^2$ (theory $96.7\epsilon^2$)
standard deviation: $2.2\epsilon^2$

Variance: $55.1\epsilon^2$ (theory $53.4\epsilon^2$)
standard deviation $1.2\epsilon^2$

Variance: $34.9\epsilon^2$ (theory $34.0\epsilon^2$)
standard deviation $0.8\epsilon^2$

One sample of $Z = \sum_{i=1}^{N} \Delta_{(\ell),i}^2/N \sim \mathrm{Var}(\Delta_{(\ell)}) + \frac{\sigma}{\sqrt{N}}\mathcal{N}(0,1)$ $\sigma^2 = \mathrm{Var}(\Delta_{(\ell)}^2)$

Validation

Model Fixed-Float Addition

Validation

Model Fixed-Float Addition

- Matrix multiplication is a common target for acceleration

Validation

Model Fixed-Float Addition

- Matrix multiplication is a common target for acceleration
- Summation is often a critical path if pairwise rounding is needed

Validation

Model Fixed-Float Addition

- Matrix multiplication is a common target for acceleration
- Summation is often a critical path if pairwise rounding is needed
- Hybrid fixed-float summation is an obvious alternative

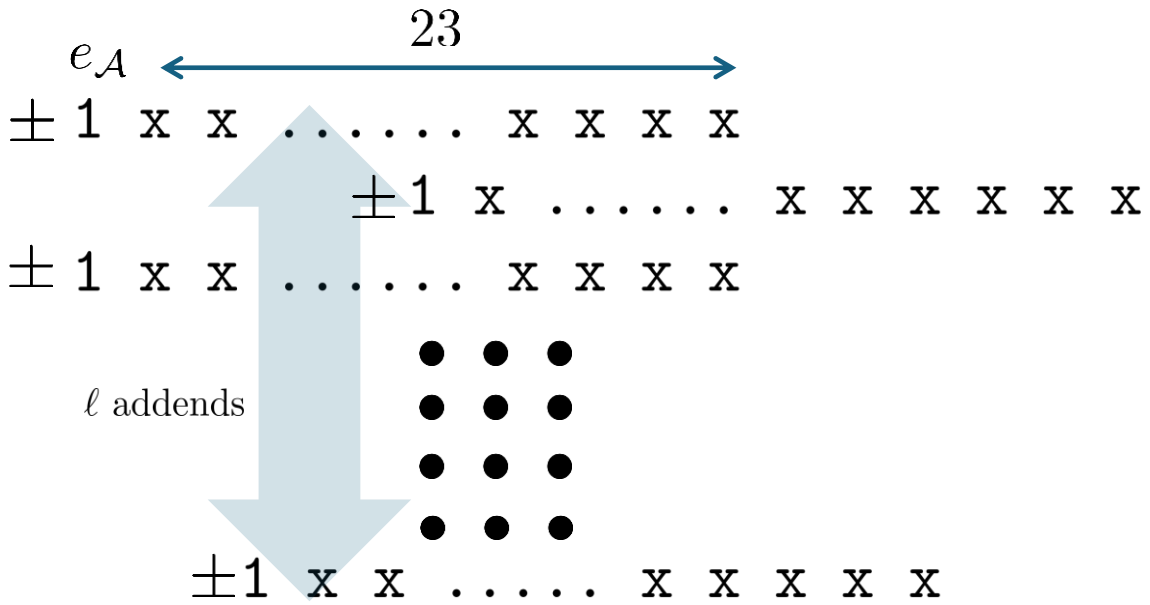
Validation

Model Fixed-Float Addition

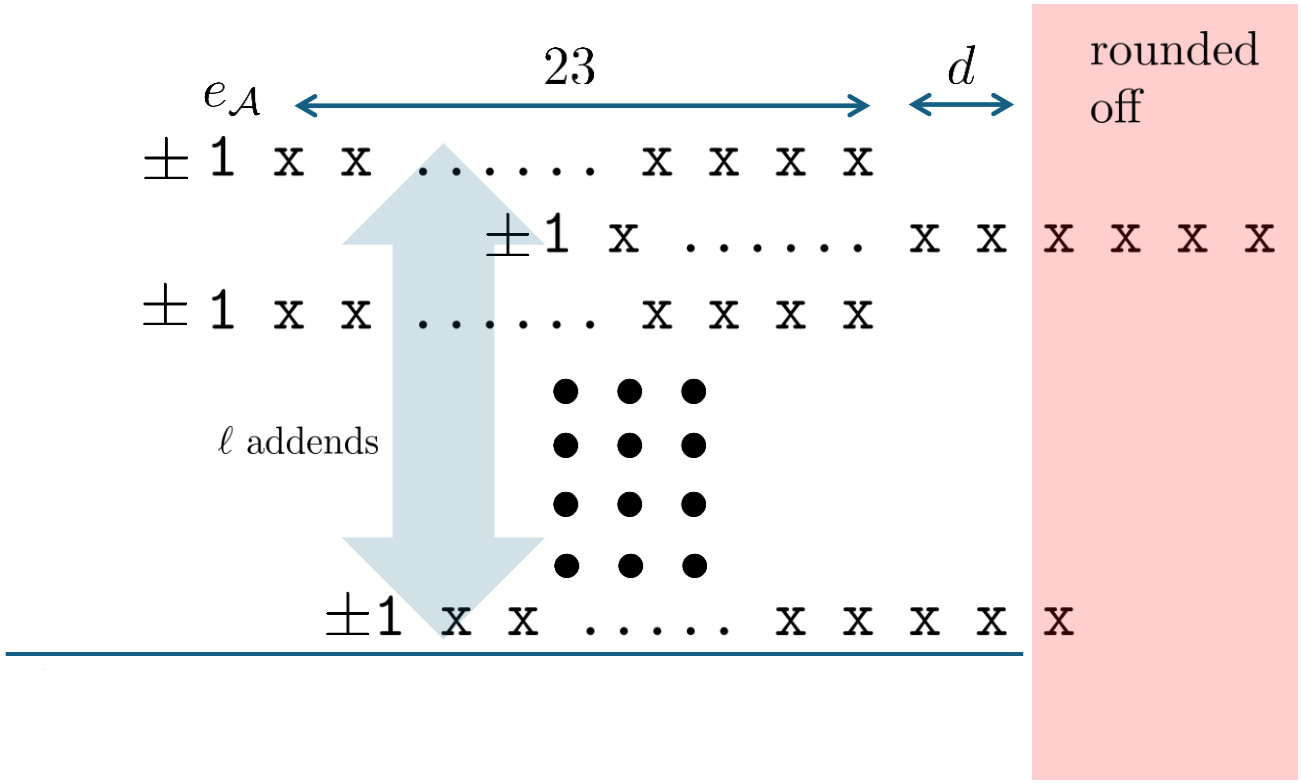
- Matrix multiplication is a common target for acceleration
- Summation is often a critical path if pairwise rounding is needed
- Hybrid fixed-float summation is an obvious alternative
- A hypothetical FP32 architecture:
  - For ℓ input fp numbers, align with the largest exponent
  - Smaller exponent values are right shifted
  - Bits beyond $23 + d$ fractional values are rounded off
  - Sum the rounded values exactly; then round to FP32

Validation

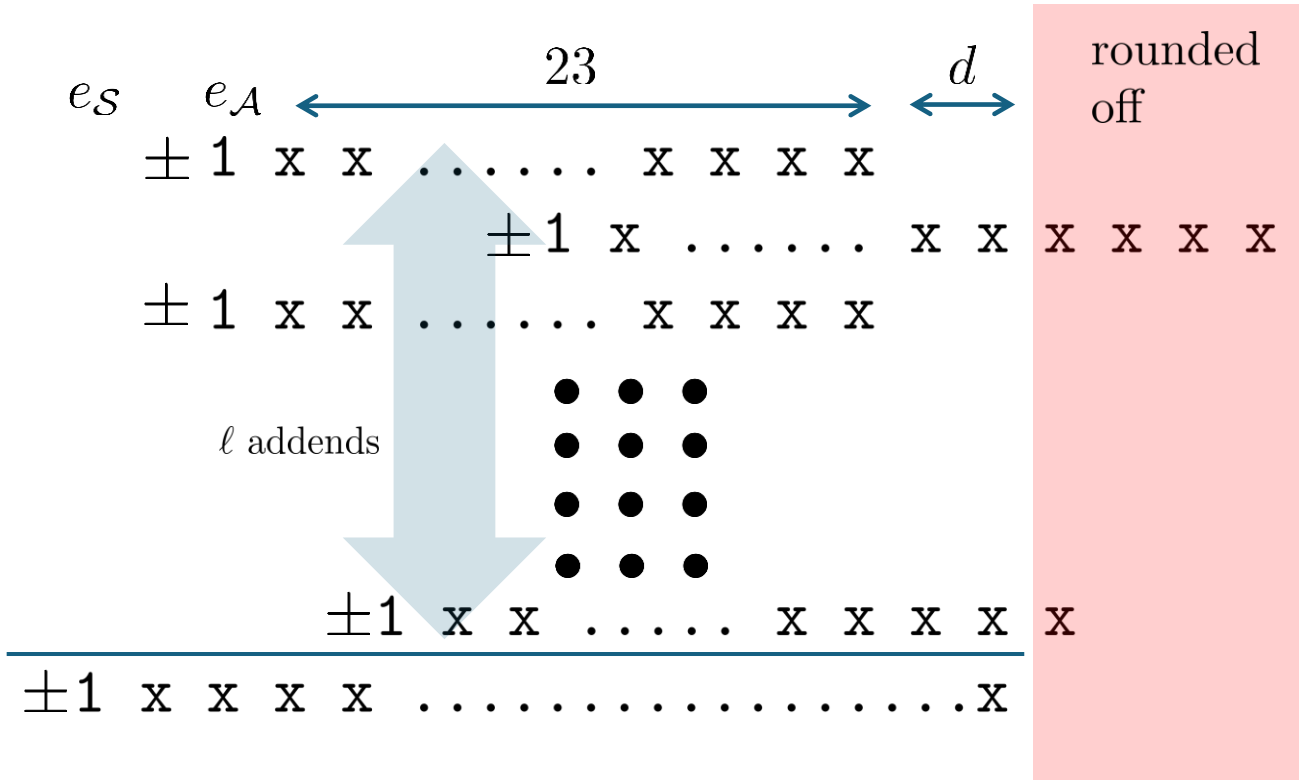
Model Fixed-Float Addition



Validation

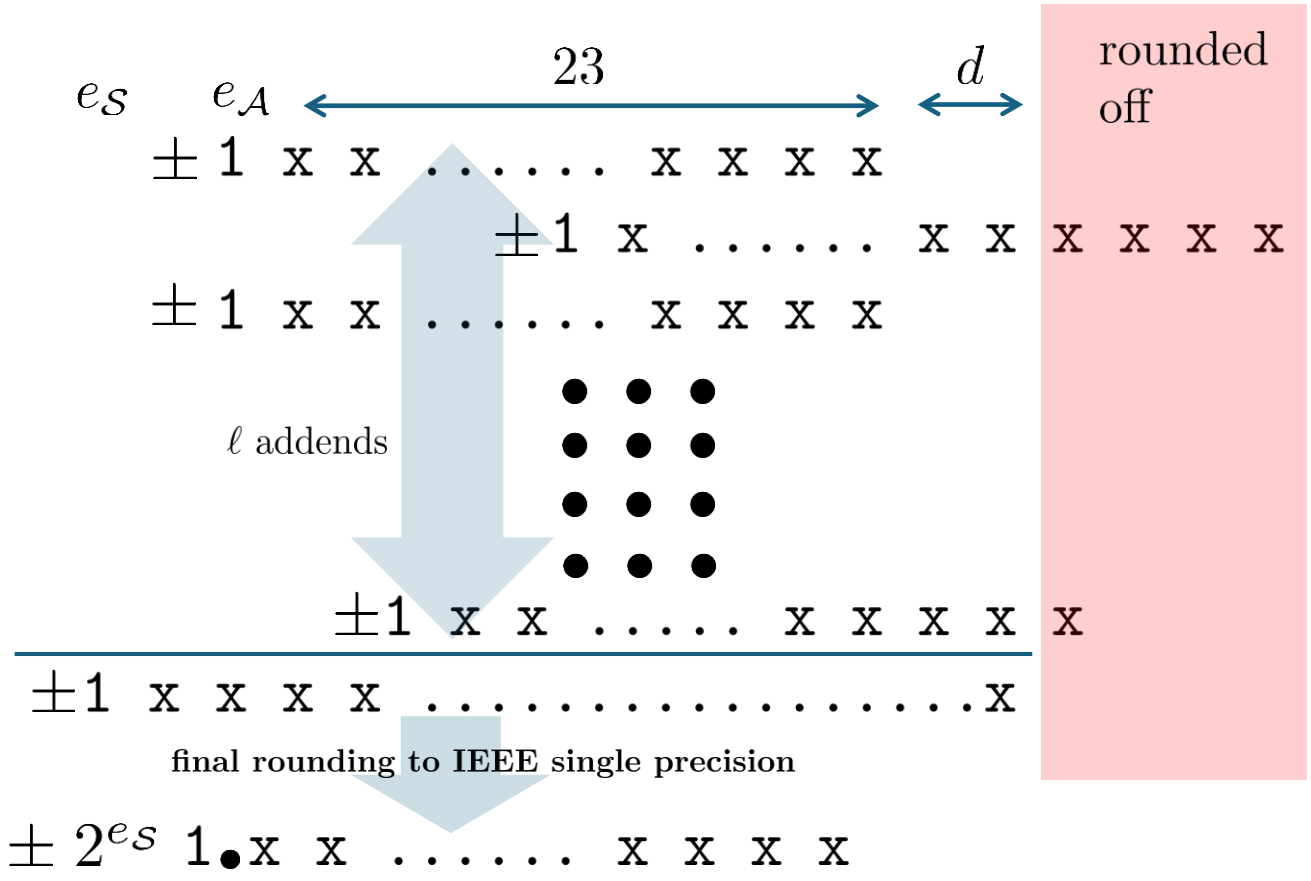

Model Fixed-Float Addition
23
$e_{\mathcal{A}}$
d
rounded off
$\pm$1 x x ..... x x x x
$\pm$1 x ..... x x x x x x
$\pm$1 x x ..... x x x x
ℓ addends
$\pm$1 x x ..... x x x x x

Validation


Model Fixed-Float Addition
e_S
e_A
23
d
rounded off
± 1 x x ..... x x x x
± 1 x ..... x x x x x x
± 1 x x ..... x x x x
ℓ addends
± 1 x x ..... x x x x x
± 1 x x x x x

Validation

Model Fixed-Float Addition



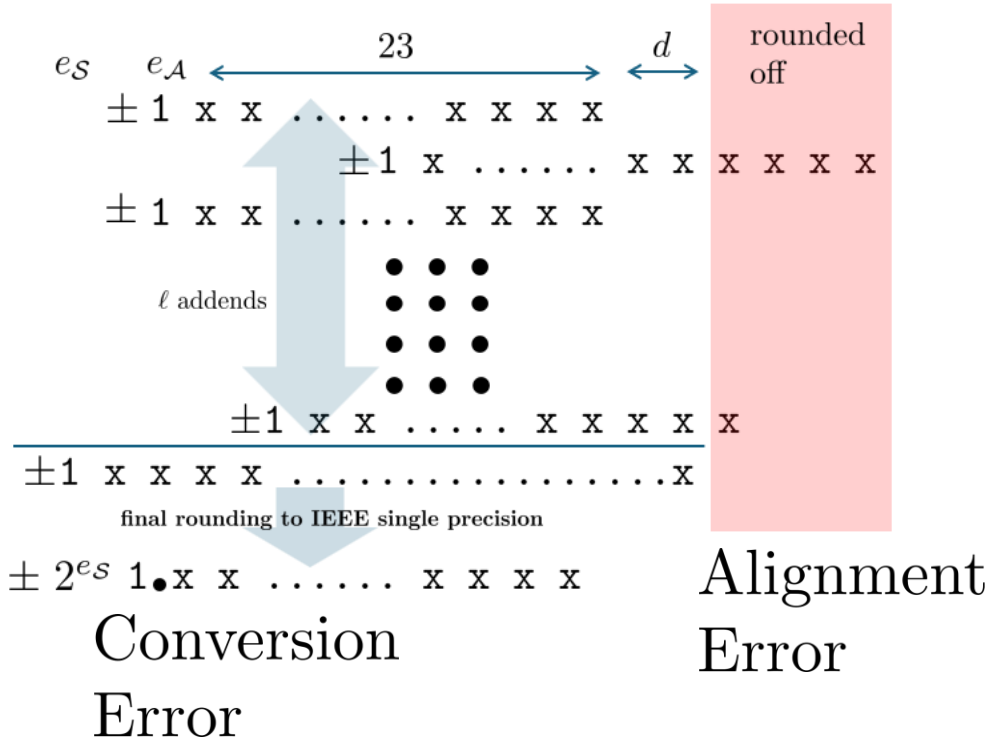
Validation

Model Fixed-Float Addition

$\Delta_{\mathcal{A}}$ denotes alignment error of one summand

v_j is variance of rounding j bits
$(v_1 = 1/8,\ v_2 = 3/32,\ \text{etc},\ \rightarrow 1/12)$

$$E(\Delta_{\mathcal{A}}^2 \mid \text{alignment expo } e_a, \text{ round } j \text{ bits})$$
$$= 2^{-2(23+d)} \qquad 2^{2e_a} \qquad v_j$$



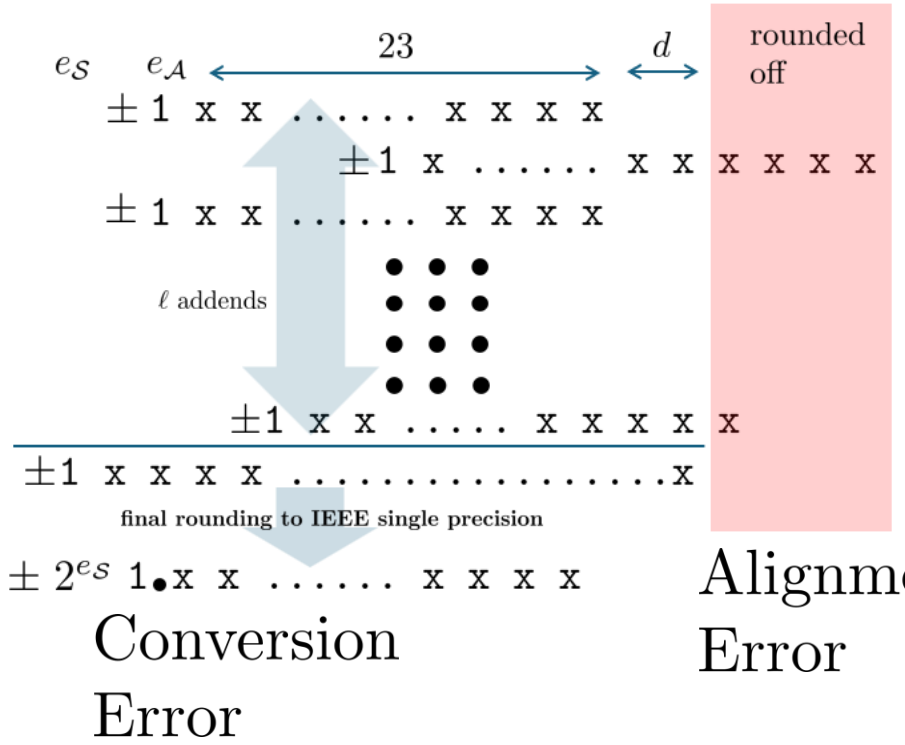
Validation

Model Fixed-Float Addition

$\Delta_{\mathcal{A}}$ denotes alignment error of one summand

v_j is variance of rounding j bits
$(v_1 = 1/8,\ v_2 = 3/32,\ \text{etc},\ \to 1/12)$

$$E(\Delta_{\mathcal{A}}^2) = 2^{-2(23+d)} \sum_{e_a} \sum_{j>1} 2^{2e_a} p_j(e_a) v_j$$



$p_j(e_a)$ is probability that

- alignment exponent of $\ell - 1$ values is e_a
- and j bits are to be rounded off

Validation

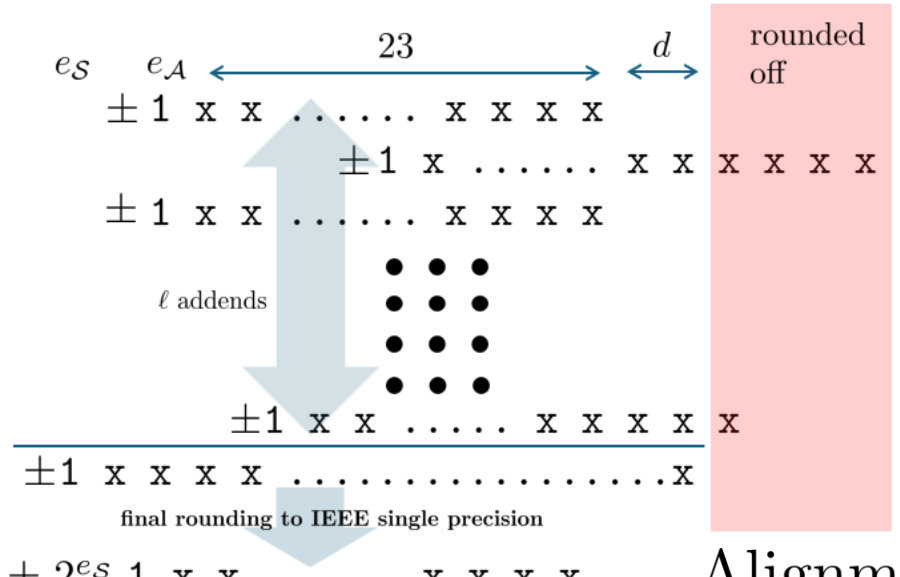
Model Fixed-Float Addition

$\Delta_{\mathcal{A}}$ denotes alignment error of one summand

v_j is variance of rounding j bits
($v_1 = 1/8$, $v_2 = 3/32$, etc, $\to 1/12$)

$$E(\Delta_{\mathcal{A}}^2) = 2^{-2(23+d)} \sum_{e_a} \sum_{j>1} 2^{2e_a} p_j(e_a) v_j$$

$$E(\Delta_{\mathcal{A}}^4) = 2^{-4(23+d)} \sum_{e_a} \sum_{j>1} 2^{4e_a} p_j(e_a) w_j$$



$p_j(e_a)$ is probability that

- alignment exponent of $\ell - 1$ values is e_a
- and j bits are to be rounded off

Validation

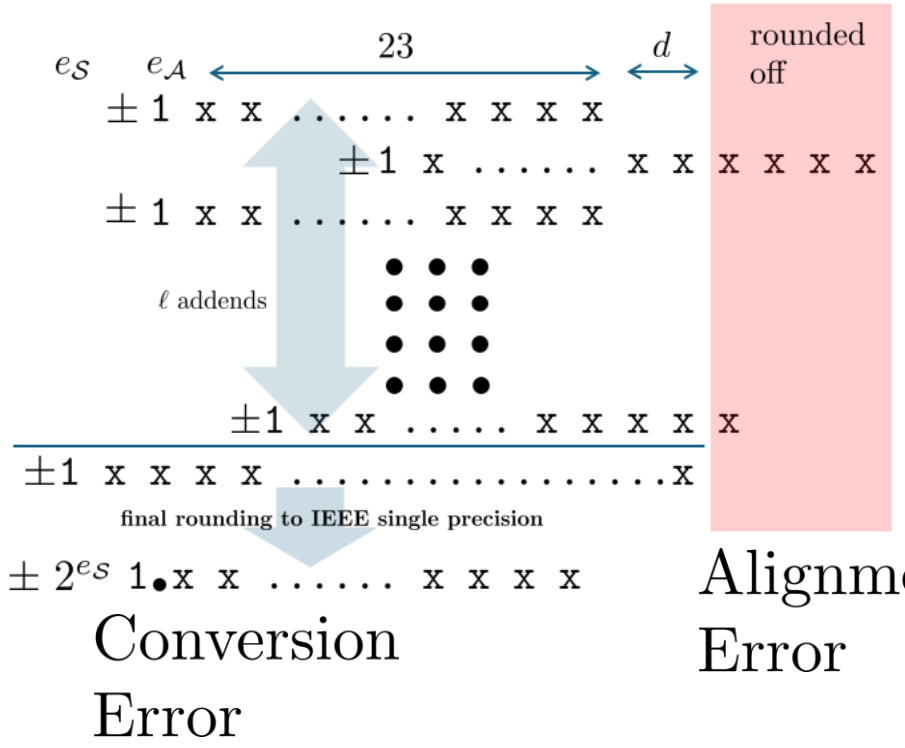
Model Fixed-Float Addition

$\Delta_{\mathcal{C}}$ denotes conversion error of final sum

v_j is variance of rounding j bits
$(v_1 = 1/8,\ v_2 = 3/32,\ \text{etc}, \rightarrow 1/12)$

$E(\Delta_{\mathcal{C}}^2 \mid \text{alignment expo } e_a, \text{ round } j \text{ bits})$

$= 2^{-2(23+d)} \qquad 2^{2(e_a+j)} \qquad v_j$



$q_j(e_a)$ is probability that
- alignment exponent of $\ell - 1$ values is e_a
- and j bits are to be rounded off

Validation

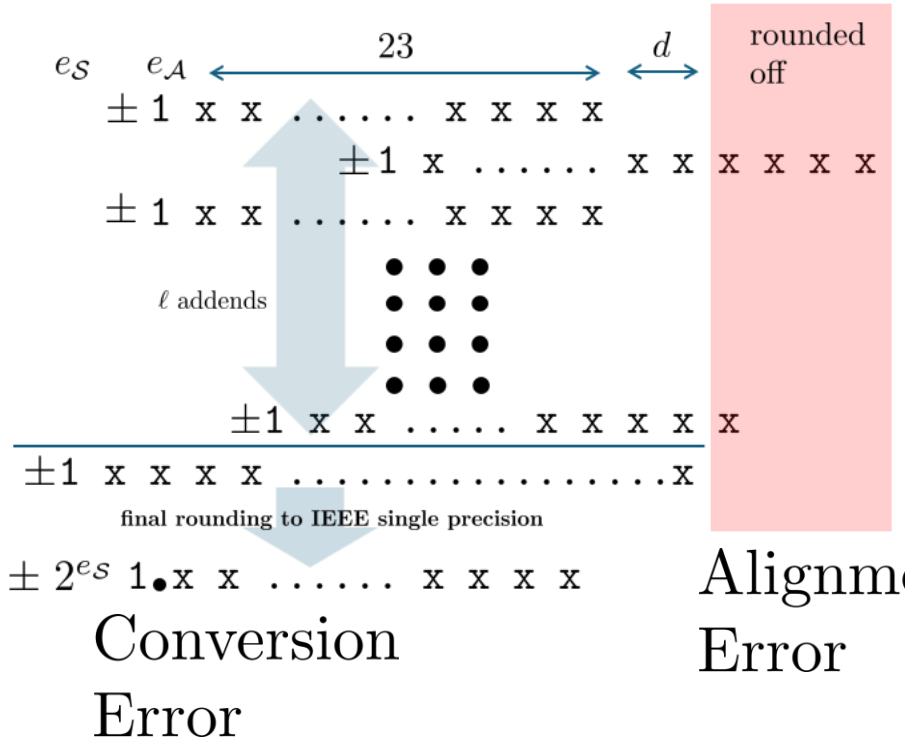
Model Fixed-Float Addition

$\Delta_{\mathcal{A}}$ denotes alignment error of one summand

v_j is variance of rounding j bits
$(v_1 = 1/8,\ v_2 = 3/32,\ \text{etc},\ \to 1/12)$

$$E(\Delta_{\mathcal{C}}^2) = 2^{-2(23+d)} \sum_{e_a} \sum_{j=1}^{\lceil \log_2(d) \rceil + d} 2^{2(e_a+j)} q_j(e_a) v_j$$

$$E(\Delta_{\mathcal{C}}^4) = 2^{-4(23+d)} \sum_{e_a} \sum_{j=1}^{\lceil \log_2(d) \rceil + d} 2^{4(e_a+j)} q_j(e_a) w_j$$



$q_j(e_a)$ is probability that

- alignment exponent of $\ell - 1$ values is e_a
- and j bits are to be rounded off

Validation

Model Fixed-Float Addition

Mean Square Error (Variance): Sampled, Model, Deviation

	$\ell = 4$			$\ell = 8$			$\ell = 16$		
Alignment	3.18e-17	3.01e-17	1.4	5.24e-17	5.31e-17	0.4	8.16e-17	8.30e-17	0.6
Conversion	2.68e-15	2.73e-15	0.7	5.27e-15	5.32e-15	0.4	1.04e-14	1.06e-14	0.5
Fixed-Float Accumulator	2.82e-15	2.85e-15	0.4	5.81e-15	5.75e-15	0.5	1.20e-14	1.19e-14	0.4

Validation

- Statistics of rounding errors are reliable metrics
- The main point of the modeling here is that we **do not** need to do it
- Statistics collected by sampling on reference numerical kernels suffices

Applications

- Set error thresholds that are more theoretically supported
- Numerical diagnostics
- Design explorations