MARCH 05, 2026

COMPENSATED SUMMATION ALGORITHMS

LONGFEI GAO

U.S. DEPARTMENT of ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

Argonne NATIONAL LABORATORY | Argonne Leadership Computing Facility

Aurora

- 10624 nodes $\times$ [2 CPUs , 6 GPUs]

Modern hardware

Where are the silicons dedicated to:

	Blackwell 200	Blackwell 300
FP8	4.5/9 petaFLOPS	4.5/9 petaFLOPS
FP16/BF16	2.2/4.5 petaFLOPS	2.2/4.5 petaFLOPS
FP32	75 teraFLOPS	75 teraFLOPS
FP64	37 teraFLOPS	1.2 teraFLOPS
memory	Up to 192 GB	270 GB
bandwidth	7.7 TB/s	7.7 TB/s

NVIDIA Blackwell Architecture Technical Brief (Built for the Age of AI Reasoning)
https://resources.nvidia.com/en-us-blackwell-architecture (accessed 2026-02-17)

Modern hardware

Where are the silicons dedicated to:

Reduced precision matrix multipliers.

$$\begin{array}{c} \text{Matrix multiplication} \\ \Downarrow \\ \text{Inner product} \\ \Downarrow \\ \text{Summation} \end{array} \qquad \text{(FMA)}$$

FMA: fused multiply add.
MMA: matrix multiply and accumulate (NVIDIA Tensor Cores).

System testing

"Anticipatory" errors:

- Roundoff errors
- Algebraic errors
- Discretization errors

"Inadvertent" errors:

- Hardware defects
- Software bugs
- Random bitflips

In designing validation tests, we want to suppress the effect of "anticipatory" errors as much as possible (to keep the "noise" level down) so that we can better detect the "inadvertent" ones.

An example

Table 1: Relative errors for calculating $s = \sum_{i=1}^{n} x_i$ using various algorithms at FP64.

	no compensation	6op	double 6op
2^2	3.0810E-17	0.0000E+00	0.0000E+00
2^4	4.7274E-53	1.3203E-69	1.3203E-69
2^6	4.4697E-17	5.7065E-20	9.3778E-34
2^8	6.9736E-17	3.9433E-18	1.1858E-32
2^{10}	1.6921E-16	2.1145E-18	3.7625E-33
2^{12}	1.6312E-15	2.2531E-17	4.7920E-32
2^{14}	2.2156E-16	1.5226E-16	2.9948E-33
2^{16}	8.2921E-16	4.3061E-18	2.5255E-32
2^{18}	1.4821E-15	1.6615E-17	1.2419E-31
2^{20}	8.4132E-15	2.7501E-17	1.3656E-30

Recursive summation with no compensation

```
s = 0;
for i = 1:n
    s = s + x[i];
end
```

Algorithm 1: Recursive summation.

Recursive summation with 6op compensation

```
s = 0;
e = 0;
for i = 1:n
      y = e + x[i];
      [s,e] = add_6op(s,y);
end
```

Algorithm 2: Recursive summation with 6op compensation.

Recursive summation with *double* 6op compensation

```
s = 0;
e = 0;
for i = 1:n
    [t,v] = add_6op(s,x[i]);
        w   = e + v;
    [s,e] = add_6op(t,w);
end
```

Algorithm 3: Recursive summation with *double* 6op compensation.

add_6op()

Theorem 1 (6op EFT over addition in Dekker's system)

Under the *as if* rule† and nearest rounding, $\forall x, y \in \mathbb{D}$, the following formulas achieve error free transformation (EFT) $z + zz = x + y$:

$$
\begin{aligned}
z &= fl(x+y);\\
w &= fl(z-x);\\
z1 &= fl(y-w);\\
v &= fl(w-z);\\
z2 &= fl(x+v);\\
zz &= fl(z1+z2).
\end{aligned}
$$

† The *as if* rule: The floating point arithmetic operation under consideration are performed *as if* an intermediate result correct to infinite precision and with unbounded range has been produced first, and then rounded to fit the destination format.

Dekker's system

$$\mathbb{D} = \left\{ x \,|\, x = m(x) \times \beta^{e(x)},\ |m(x)| < \beta^t,\ E_{\min} \le e(x) \le E_{\max} \right\}$$

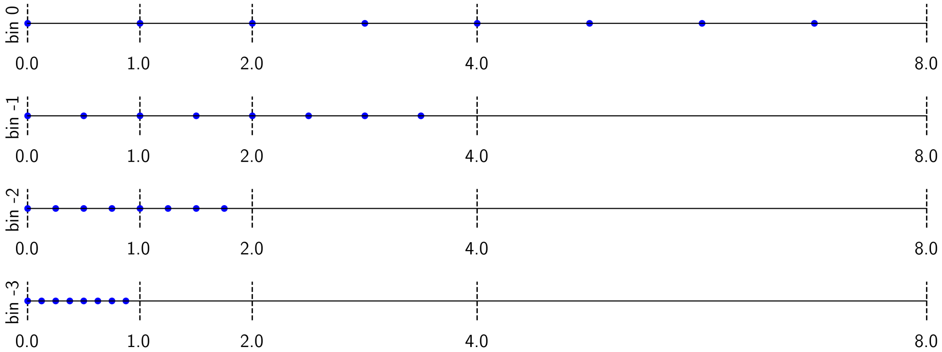


Figure 1: Representations in Dekker's system with $\beta = 2$, $t = 3$, $E_{\min} = -3$, and $E_{\max} = 0$.

IEEE system

$$\mathbb{F} = \left\{ x \mid x = m(x) \times \beta^{e(x)},\ \beta^{t-1} \leq |m(x)| < \beta^{t},\ E_{\min} \leq e(x) \leq E_{\max} \right\}$$

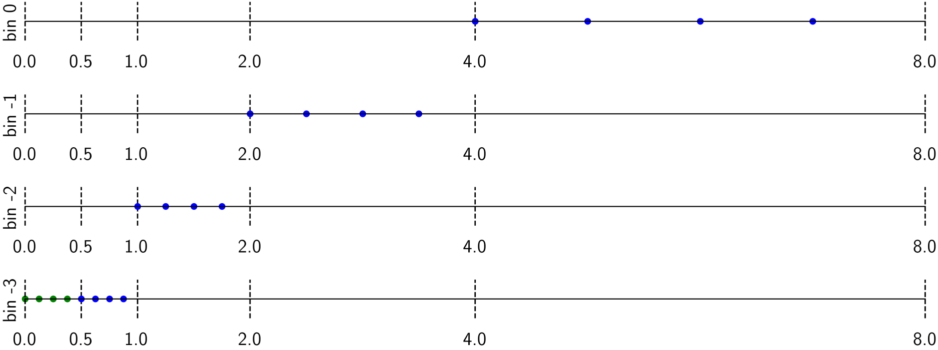


Figure 2: Representations in the proto-IEEE system with $\beta = 2$, $t = 3$, $E_{\min} = -3$, and $E_{\max} = 0$.

Proof of Theorem 1

Proof. The case when x and y admit representations with $e(x) \geq e(y)$ has already been covered by the proof of Theorem 3.1 since it is easy to show that $z2 = 0$ following $w = z - x$ in this case. Therefore, we limit our discussion below to the complementary case, i.e., when x and y *do not* admit representations such that $e(x) \geq e(y)$.

In this scenario, aside from the obvious condition $d = e(x) - e(y) < 0$, we also have that y cannot admit a representation with the exponent $E_{\min}$ since, otherwise, any representation of x would lead to $e(x) \geq e(y)$ for this representation of y. Consequently, y admits a normal representation in $\mathbb{D}$. We work with this normal representation of y in the following. In other words, when writing $y = m(y) \times \beta^{e(y)}$ below, the conditions $e(y) > E_{\min}$ and $\beta^{t-1} \leq |m(y)| < \beta^t$ are implied. These conditions appear in Lemma 3.3 and Lemma 3.4 invoked in this proof.

The proof below consists of three parts: first, we show that $z1 = y - w$ under the above conditions; second, we show that $w - z2 = z - x$ by invoking Theorem 3.1; third, we show that $zz = z1 + z2$.

We proceed to show that $z1 = y - w$ still holds by separating the discussion into four cases based on the signs of f and g, similar to that in the proof of Theorem 3.1. We assume that overflow does not happen when forming z for the four cases below. The special scenario of overflow is addressed immediately after the discussion of these cases.

Case 1: if $f \geq 0$ and $g > 0$, we again have $\zeta \in \mathbb{Z}$ as in Case 1 of Theorem 3.1, where ζ is defined in (21), and thus, $z = x + y$. This leads to $w = z - x = y$ and $z1 = fl(y - w) = 0 = y - w$ as desired. Additionally, we also have $v = -x$, $z2 = 0$, and $zz = 0$ following $w = z - x$.

Case 2: if $f \geq 0$ and $g \leq 0$, we have $d = f - g \geq 0$, which contradicts the assumption $d < 0$. Thus, this is a null case.

Case 3: if $f < 0$ and $g \geq 0$, recall that $z = m(z) \times \beta^{e(z)} = round(\zeta) \times \beta^{e(z)}$, where ζ is defined in (21), we have

$$\begin{aligned} z - x &= \left(m(z) - m(x) \cdot \beta^f\right) \times \beta^{e(z)} && \text{(28a)}\\ &= \left(m(z) - m(x) \cdot \beta^f - m(y) \cdot \beta^g + m(y) \cdot \beta^g\right) \times \beta^{e(z)} && \text{(28b)}\\ &= y + \left(m(z) - m(x) \cdot \beta^f - m(y) \cdot \beta^g\right) \times \beta^{e(z)} && \text{(28c)}\\ &\in \left[y - \tfrac{1}{2} \times \beta^{e(z)}, y + \tfrac{1}{2} \times \beta^{e(z)}\right] \subseteq \left[y - \tfrac{1}{2} \times \beta^{e(y)}, y + \tfrac{1}{2} \times \beta^{e(y)}\right]. && \text{(28d)} \end{aligned}$$

From Lemma 3.3 presented after this theorem, we know that $y - fl(z - x) = y - w$ can only take values $\pm 1 \times \beta^{e(y)-1}$, 0, or $\pm 1 \times \beta^{e(y)}$, all of which are in $\mathbb{D}$. Therefore, we have $z1 = y - w$ as desired.

Case 4: if $f < 0$ and $g < 0$, we have that z admits a representation with $e(z) \leq e(y) + 1$ following Lemma 2.12. Working with a representation of z such that $e(z) \leq e(y) + 1$ holds, the condition that $g < 0$ further implies that $e(z) = e(y) + 1$. Similar to (28), we have in this case

$$z - x \in \left[y - \tfrac{1}{2} \times \beta^{e(z)}, y + \tfrac{1}{2} \times \beta^{e(z)}\right] = \left[y - 1 \times \beta^{e(y)}, y + 1 \times \beta^{e(y)}\right], \tag{29}$$

where $\beta = 2$ is used for the last equality. When $t > 1$, we can invoke Lemma 3.4 presented after this theorem to arrive at the desired result $z1 = y - w$ as in Case 3. The special situation $t = 1$ is addressed in Remark 6 following this theorem.

We now address the special scenario of overflow when forming z. Notice that overflow cannot possibly happen in Cases 1 and 2 above because we have $e(z) < e(y)$ under the combination of assumptions $d < 0$ and $f \geq 0$. Therefore, we need only to concentrate on Cases 3 and 4 above, where $f < 0$. Moreover, since x and y must have the same sign for z to overflow, it is sufficient to address only the case where both x and y are positive. Because overflow leads to clipping in Dekker's system, we have $z = (\beta^t - 1) \times \beta^{E_{\max}}$ and, therefore,

$$z - x = \left((\beta^t - 1) - m(x) \cdot \beta^f\right) \times \beta^{E_{\max}} \geq \left((\beta^t - 1) - (\beta^t - 1) \cdot \beta^{-1}\right) \times \beta^{E_{\max}} = (\beta^t - 1) \times \beta^{E_{\max}-1},$$

where $f \leq -1$ is used for the inequality and $\beta = 2$ is used for the last equality. Moreover, because of overflow, we also have $x + y > z$, i.e., $z - x < y$. Thus, by monotonicity, we have

$$(\beta^t - 1) \times \beta^{E_{\max}-1} \leq w \leq y. \tag{30}$$

Because of (30), we can write

$$\begin{aligned} y &= \tilde{m}(y) \times \beta^{E_{\max}-1}; && \text{(31a)}\\ w &= \tilde{m}(w) \times \beta^{E_{\max}-1}, && \text{(31b)} \end{aligned}$$

where $\tilde{m}(y) \in \mathbb{Z}$ and $\tilde{m}(w) \in \mathbb{Z}$. Notice that we have used $\tilde{m}(\cdot)$ instead of $m(\cdot)$ for the mantissas in (31a) and (31b) because they are not necessarily valid representations in $\mathbb{D}$. Moreover, since $\tilde{m}(y) \leq (\beta^t - 1) \cdot \beta$ and $\tilde{m}(w) \geq (\beta^t - 1)$, we can write

$$y - w = (\tilde{m}(y) - \tilde{m}(w)) \times \beta^{E_{\max}-1}, \tag{32}$$

where $\tilde{m}(y) - \tilde{m}(w) \in \mathbb{Z}$ and $0 \leq \tilde{m}(y) - \tilde{m}(w) \leq \beta^t - 1$. Thus, $y - w \in \mathbb{D}$ with (32) being a valid representation, which leads to $z1 = y - w$ as desired. This concludes the discussion for the special scenario of overflow when forming z.

Now that we completed the demonstration of $z1 = y - w$, we proceed to show that $w - z2 = z - x$. When $f \geq 0$, from the discussion above, Case 1 is the only nontrivial case under the conditions of the theorem, for which $w - z2 = z - x$ holds because $w = z - x$ and $z2 = 0$. When $f < 0$, i.e., $e(x) < e(z)$, following Theorem (3.1), operations (27b), (27d), and (27e) form a 3op EFT over addition (after adjusting for the signs), which leads to $w - z2 = z - x$. Combined with $z1 = y - w$ obtained earlier, this leads to

$$z + z1 + z2 = x + y. \tag{33}$$

It remains to show that $zz = z1 + z2$, i.e., (27f) also commits no rounding error. For all non-empty cases discussed above, including the special scenario of overflow when forming z and the special systems with $t = 1$, we have that $z1 = y - w$ admits a representation in $\mathbb{D}$ with an exponent no less than $e(y) - 1$. Since $d = e(x) - e(y) < 0$, it follows that $z1$ can be expressed as $\tilde{m}(z1) \times \beta^{e(x)}$ where $\tilde{m}(z1) \in \mathbb{Z}$. Moreover, from the proof of Theorem (3.1), we have that $z2$ can be expressed as $m(z2) \times \beta^{e(x)}$ where $m(z2) \in \mathbb{Z}$ and $|m(z2)| < M$. Therefore, $z1 + z2$ can be expressed as an integer times $\beta^{e(x)}$ as well, i.e., $z1 + z2 = \left(\tilde{m}(z1) + m(z2)\right) \times \beta^{e(x)}$. From (33), we have

$$|z1 + z2| = |x + y - z| \leq |x + y - y| = |x|$$

following the nearest rounding property, which leads to $|\tilde{m}(z1) + m(z2)| \leq |m(x)| < M$. Therefore, $z1 + z2 \in \mathbb{D}$ and $zz = z1 + z2$ follows.

End of proof for Theorem 3.2.

***Remark* 6** (Special scenario: $t = 1$). We use this remark to address the special scenario when $t = 1$ in Case 4 of Theorem 3.2. When $t = 1$, the mantissa can only take value 0 or ± 1. If $z = 0$, z obviously admits a representation with exponent $E_{\min}$ and, hence, we can invoke Lemma 2.15 to arrive at $z = x + y$. If either x or y is 0, $z = x + y$ also holds since $x, y \in \mathbb{D}$. Following the same argument as in Case 1 of Theorem 3.2, we arrive at $zz = 0$ and $z + zz = x + y$ subsequently.

Given the above observation, we assume that x, y, and z all have nonzero mantissa below. Recall the conditions $f = e(x) - e(z) < 0$ and $g = e(y) - e(z) < 0$ for Case 4 of Theorem 3.2, combined with $t = 1$, we have that $|x| < |z|$ and $|y| < |z|$, which imply that x and y must have the same sign, and so does z. Therefore, it suffices to consider the case where x, y, and z all have mantissa 1.

Recall that z admits a representation with $e(z) = e(y) + 1$ from Case 4 of Theorem 3.2, we can write $y = 1 \times \beta^{e(y)}$ and $z = 1 \times \beta^{e(y)+1}$. Moreover, following the condition $d = e(x) - e(y) < 0$, it is necessary that $e(x) = e(y) - 1$ and $x = 1 \times \beta^{e(y)-1}$ because, otherwise, i.e., if $e(x) < e(y) - 1$, $x + y$ would round to y instead of z.

Therefore, we have $z - x = (1 + {}^1\!/\!{}_2) \times \beta^{e(y)}$, which means that $w = fl(z - x)$ can be either $1 \times \beta^{e(y)}$ or $1 \times \beta^{e(y)+1}$, depending on the tie-breaking rule. Therefore, $y - w$ can only take value 0 or $-1 \times \beta^{e(y)}$, both of which are in $\mathbb{D}$. Hence, we arrive at the desired result $z1 = y - w$.

***End of Remark* 6.**

The following lemma is invoked for Case 3 in the proof of Theorem 3.2.

Lemma 3.3. *With $\beta = 2$ and nearest rounding, given $y = m(y) \times \beta^{e(y)} \in \mathbb{D}$ with $\beta^{t-1} \leq |m(y)| < \beta^t$ and $e(y) > E_{\min}$, the rounding outcome of a real number $\tilde{w} \in \left[y - \frac{1}{2} \times \beta^{e(y)}, y + \frac{1}{2} \times \beta^{e(y)}\right]$ admits a normal representation with exponent $e(y) - 1$, $e(y)$, or $e(y) + 1$. Moreover, the difference $y - fl(\omega)$ can take values $\pm 1 \times \beta^{e(y)-1}$, 0, or $\pm 1 \times \beta^{e(y)}$ only.*

Proof. We have $y \neq 0$ since $|m(y)| \geq \beta^{t-1}$. Moreover, it suffices to consider only the case $y > 0$ because of symmetry.

We first address the special case $t = 1$, where the mantissa can only take value 1. We can relax the interval to $\left[y - \frac{1}{2} \times \beta^{e(y)}, y + 1 \times \beta^{e(y)}\right]$, which becomes $\left[1 \times \beta^{e(y)-1}, 1 \times \beta^{e(y)+1}\right]$ in this case. Since both endpoints of the above interval are in $\mathbb{D}$, by Lemma 2.8 (Monotonicity), we have that $\tilde{w}$ can only round to one of the following three floating point numbers:

$$\left[1 \times \beta^{e(y)-1}, 1 \times \beta^{e(y)+1}\right] \cap \mathbb{D} = \left\{1 \times \beta^{e(y)-1}, 1 \times \beta^{e(y)}, 1 \times \beta^{e(y)+1}\right\}. \tag{34}$$

Consequently, the difference $y - fl(\tilde{w})$ can only take values

$$\left\{1 \times \beta^{e(y)-1},\ 0,\ \text{and}\ -1 \times \beta^{e(y)}\right\}. \tag{35}$$

We note here that the last member of (34), i.e., $1 \times \beta^{e(y)+1}$, may overflow. In that case, $\tilde{w}$ can only round to the first two members of (34) in Dekker's system.

With the above special treatment for $t = 1$, we assume $t > 1$ and separate the discussion below based on the size of $m(y)$.

Case 1: if $m(y) = \beta^{t-1}$, we can relax the interval to $\left[y - \frac{1}{2} \times \beta^{e(y)}, y + 1 \times \beta^{e(y)}\right]$, which becomes $\left[(\beta^t - 1) \times \beta^{e(y)-1}, (\beta^{t-1} + 1) \times \beta^{e(y)}\right]$ in this case, with both endpoints in $\mathbb{D}$. By monotonicity, $\tilde{w}$ can only round to one of the following three floating point numbers:

$$\left[(\beta^t - 1) \times \beta^{e(y)-1}, (\beta^{t-1} + 1) \times \beta^{e(y)}\right] \cap \mathbb{D} = \left\{(\beta^t - 1) \times \beta^{e(y)-1}, \beta^{t-1} \times \beta^{e(y)}, (\beta^{t-1} + 1) \times \beta^{e(y)}\right\}. \tag{36}$$

Consequently, the difference $y - fl(\tilde{w})$ can only take values

$$\left\{1 \times \beta^{e(y)-1},\ 0,\ \text{and}\ -1 \times \beta^{e(y)}\right\}. \tag{37}$$

Case 2: if $\beta^{t-1} < m(y) < \beta^t - 1$, we can relax the interval to $\left[y - 1 \times \beta^{e(y)}, y + 1 \times \beta^{e(y)}\right]$, which becomes $\left[(m(y) - 1) \times \beta^{e(y)}, (m(y) + 1) \times \beta^{e(y)}\right]$ in this case, with both endpoints in $\mathbb{D}$. By monotonicity, $\tilde{w}$ can only round to one of the following three floating point numbers:

$$\left[(m(y) - 1) \times \beta^{e(y)}, (m(y) + 1) \times \beta^{e(y)}\right] \cap \mathbb{D} = \left\{(m(y) - 1) \times \beta^{e(y)}, m(y) \times \beta^{e(y)}, (m(y) + 1) \times \beta^{e(y)}\right\}. \tag{38}$$

Consequently, the difference $y - fl(\tilde{w})$ can only take values

$$\left\{1 \times \beta^{e(y)},\ 0,\ \text{and}\ -1 \times \beta^{e(y)}\right\}. \tag{39}$$

We note here that if $t = 1$ or 2, we would have $\beta^{t-1} + 1 \geq \beta^t - 1$, leading to a null case here.

Case 3: if $m(y) = \beta^t - 1$, we can relax the interval to $\left[y - 1 \times \beta^{e(y)}, y + 1 \times \beta^{e(y)}\right]$, which becomes $\left[(\beta^t - 2) \times \beta^{e(y)}, \beta^{t-1} \times \beta^{e(y)+1}\right]$ in this case, with both endpoints in $\mathbb{D}$. By monotonicity, $\tilde{w}$ can only round to one of the following three floating point numbers:

$$\left[(\beta^t - 2) \times \beta^{e(y)}, (\beta^{t-1}) \times \beta^{e(y)+1}\right] \cap \mathbb{D} = \left\{(\beta^t - 2) \times \beta^{e(y)}, (\beta^t - 1) \times \beta^{e(y)}, \beta^{t-1} \times \beta^{e(y)+1}\right\}. \tag{40}$$

Consequently, the difference $y - fl(\tilde{w})$ can only take values

$$\left\{1 \times \beta^{e(y)},\ 0,\ \text{and}\ -1 \times \beta^{e(y)}\right\}. \tag{41}$$

We note here that the last member of (40), i.e., $\beta^{t-1} \times \beta^{e(y)+1}$, may overflow. In that case, $\tilde{w}$ can only round to the first two members of (40) in Dekker's system.

End of proof for Lemma 3.3.

The following lemma is invoked for Case 4 in the proof of Theorem 3.2. It relaxes the interval to a larger one but excludes the special case where $t = 1$.

Lemma 3.4. *With $\beta = 2$, $t > 1$, and nearest rounding, given $y = m(y) \times \beta^{e(y)} \in \mathbb{D}$ with $\beta^{t-1} \leq |m(y)| < \beta^t$ and $e(y) > E_{\min}$, the rounding outcome of a real number $\tilde{w} \in \left[y - 1 \times \beta^{e(y)}, y + 1 \times \beta^{e(y)}\right]$ admits a normal representation with exponent $e(y) - 1$, $e(y)$, or $e(y) + 1$. Moreover, the difference $y - fl(\tilde{w})$ can take values $\pm 1 \times \beta^{e(y)-1}$, 0, or $\pm 1 \times \beta^{e(y)}$ only.*

Proof. The proof mostly follows that of Lemma 3.3. As in Lemma 3.3, it suffices to consider only the case $y > 0$.

Case 1: if $m(y) = \beta^{t-1}$, the interval becomes $\left[(\beta^t - \beta) \times \beta^{e(y)-1}, (\beta^{t-1} + 1) \times \beta^{e(y)}\right]$, with both endpoints in $\mathbb{D}$. By monotonicity, $\tilde{w}$ can only round to one of the following $\beta + 2$ (i.e., 4 since $\beta = 2$) floating point numbers between the two endpoints:

$$\left\{(\beta^t - 2) \times \beta^{e(y)-1},\ (\beta^t - 1) \times \beta^{e(y)-1},\ \beta^{t-1} \times \beta^{e(y)},\ (\beta^{t-1} + 1) \times \beta^{e(y)}\right\}. \tag{42}$$

Consequently, the difference $y - fl(\tilde{w})$ can only take values

$$\left\{2 \times \beta^{e(y)-1} = 1 \times \beta^{e(y)},\ 1 \times \beta^{e(y)-1},\ 0,\ \text{and}\ -1 \times \beta^{e(y)}\right\}. \tag{43}$$

Case 2: if $\beta^{t-1} < m(y) < \beta^t - 1$, this case has already been covered in the proof of Case 2 of Lemma 3.3.

Case 3: if $m(y) = \beta^t - 1$, this case has already been covered in the proof of Case 3 of Lemma 3.3.

End of proof for Lemma 3.4.

Before proceeding, we introduce the notation ϵ for the quantity *unit round-off*. Given a floating point number system with t-digit mantissa, ϵ is defined as $\epsilon = \frac{1}{2} \cdot \beta^{1-t}$ under nearest rounding. The following lemma regarding ϵ can be established.

Lemma 3.5. *For $\tilde{z} \in \mathbb{R}$ that rounds to $z = m(z) \times \beta^{e(z)}$ in $\mathbb{D}$, if $|\tilde{z}| \leq (M - 1) \times \beta^{E\max}$, i.e., $\tilde{z}$ does not overflow, we have $|z - \tilde{z}| \leq \frac{1}{2} \times \beta^{e(z)}$. Moreover, when z is not a subnormal number, we also have $|z - \tilde{z}| \leq \epsilon|z|$.*

Proof. Expressing $\tilde{z}$ as $\tilde{z} = \zeta \times \beta^{e(z)}$, we have $m(z) = round(\zeta)$ following Lemma 2.11 and, therefore, $|m(z) - \zeta| \leq \frac{1}{2}$. It follows that $|z - \tilde{z}| = |m(z) - \zeta| \times \beta^{e(z)} \leq \frac{1}{2} \times \beta^{e(z)}$.

Notice that the result above holds for any valid representation of z. When z is not a subnormal number, working with its normal representation $m(z) \times \beta^{e(z)}$ where $|m(z)| \geq \beta^{t-1}$, we have $|z - \tilde{z}| \leq \frac{1}{2} \times \beta^{e(z)} = \frac{1}{2} \cdot \beta^{1-t} \cdot \beta^{t-1} \times \beta^{e(z)} \leq \epsilon|z|$.

End of proof for Lemma 3.5.

Lemma 3.5 provides a statement regarding rounding in general. If the number to be rounded, i.e., $\tilde{z}$, is the intermediate result from an addition (or subtraction) operation, the additional requirement that *z is not a subnormal number* can be dropped from the statement, which is summarized below.

Lemma 3.6. *Given $x, y \in \mathbb{D}$ and $\tilde{z} = x + y \in \mathbb{R}$, which rounds to $z = m(z) \times \beta^{e(z)}$ in $\mathbb{D}$, if $|\tilde{z}| \leq (M - 1) \times \beta^{E\max}$, i.e., $\tilde{z}$ does not overflow, we have $|z - \tilde{z}| \leq \frac{1}{2} \times \beta^{e(z)}$ and $|z - \tilde{z}| \leq \epsilon|z|$.*

Proof. It suffices to consider only the case when z is a subnormal number. In this case, we have $z = x + y = \tilde{z}$ from Lemma 2.15 and, hence, the stated results still hold.

End of proof for Lemma 3.6.

The following corollary results from a straightforward application of Lemma 3.6 by observing that $zz = x + y - z = x + y - fl(x + y)$ when the conditions of Theorem 3.1 or Theorem 3.2 are met.

Corollary 3.7. *Under the conditions of Theorem 3.1 or the conditions of Theorem 3.2 and assuming overflow does not happen when forming z, we have $|zz| \leq \frac{1}{2} \times \beta^{e(z)}$ and $|zz| \leq \epsilon|z|$.*

Lemma 3.6 bounds the error with the *output* of the floating point addition operation. The lemma below bounds the error with the *inputs* of the addition operation.

Dekker's floating point number system and compensated summation algorithms, arXiv:2602.19452.

Error analysis for Algorithm 1

$$s_i \leftarrow s_{i-1} + x_i$$

```
s = 0;
for i = 1:n
      s = s + x[i];
end
```

$$\begin{aligned} s_i &= (1+\delta_i)(s_{i-1}+x_i) \\ &= (1+\delta_i)s_{i-1} \;+\; (1+\delta_i)x_i \quad ; \quad |\delta_i| \le \epsilon \end{aligned}$$

$$\begin{aligned} s_n &= \textstyle\sum_{k=1}^{n} \left\{ \prod_{i=k+1}^{n} (1+\delta_i) \right\} (1+\delta_k) x_k \\ &= \textstyle\sum_{k=1}^{n} (1+\theta_{n-k})(1+\delta_k) x_k \end{aligned}$$

$$\begin{aligned} |s_n - \tilde{s}| &= \left| \textstyle\sum_{k=1}^{n} (\theta_{n-k} + \delta_k + \theta_{n-k}\delta_k) x_k \right| \\ &\le \frac{(n-1)\epsilon}{1-(n-1)\epsilon} \textstyle\sum_{k=1}^{n} |x_k| + \epsilon \textstyle\sum_{k=1}^{n} |x_k| \\ &\quad + \frac{(n-1)\epsilon^2}{1-(n-1)\epsilon} \textstyle\sum_{k=1}^{n} |x_k| \end{aligned}$$

These relations remain valid with subnormal numbers.

Lemma 3.1. *If* $|\delta_i| \le u$ *and* $\rho_i = \pm 1$ *for* $i = 1\!:\!n$*, and* $nu < 1$*, then*

$$\prod_{i=1}^{n} (1+\delta_i)^{\rho_i} = 1 + \theta_n,$$

where

$$|\theta_n| \le \frac{nu}{1-nu} =: \gamma_n.$$

Accuracy and stability of numerical algorithms, 2002.

Error analysis for Algorithm 2

$$y_i \leftarrow e_{i-1} + x_i$$
$$s_i, e_i \leftarrow s_{i-1} + y_i$$

```
s = 0;
e = 0;
for i = 1:n
    y = e + x[i];
    [s,e] = add_6op(s,y);
end
```

$$\begin{aligned} y_i &= (1+\alpha_i)(e_{i-1}+x_i) \quad ; \quad |\alpha_i| \le \epsilon \\ s_i &= (1+\beta_i)(s_{i-1}+y_i) \quad ; \quad |\beta_i| \le \epsilon \\ e_i &= \quad -\beta_i \; (s_{i-1}+y_i) \end{aligned}$$

$$s_i + e_i = (1-\alpha_i\beta_{i-1})(s_{i-1}+e_{i-1}) + (1+\alpha_i)x_i$$

$$\begin{aligned} s_n + e_n &= \textstyle\sum_{k=1}^{n} \left\{ \prod_{i=k+1}^{n} (1-\alpha_i\beta_{i-1}) \right\} (1+\alpha_k)x_k \\ &= \textstyle\sum_{k=1}^{n} (1+\tilde{\theta}_{n-k})(1+\alpha_k)x_k \quad ; \quad |\tilde{\theta}_m| \le \frac{m\epsilon^2}{1-m\epsilon^2} \end{aligned}$$

$$\begin{aligned} |s_n + e_n - \tilde{s}| &= \left| \textstyle\sum_{k=1}^{n} (\tilde{\theta}_{n-k} + \alpha_k + \tilde{\theta}_{n-k}\alpha_k) x_k \right| \\ &\le \frac{(n-1)\epsilon^2}{1-(n-1)\epsilon^2} \textstyle\sum_{k=1}^{n} |x_k| + \epsilon \sum_{k=1}^{n} |x_k| + \frac{(n-1)\epsilon^3}{1-(n-1)\epsilon^2} \sum_{k=1}^{n} |x_k| \end{aligned}$$

These relations remain valid with subnormal numbers.

Error analysis for Algorithm 3

$$
\begin{aligned}
t_i, v_i &\leftarrow s_{i-1} + x_i \\
w_i &\leftarrow e_{i-1} + v_i \\
s_i, e_i &\leftarrow t_i + w_i
\end{aligned}
$$

```
s = 0;
e = 0;
for i = 1:n
      [t,v] = add_6op(s,x[i]);
          w = e + v;
      [s,e] = add_6op(t,w);
end
```

$$
\begin{aligned}
t_i &= (1+\alpha_i)(s_{i-1}+x_i) \quad &; \quad |\alpha_i| \le \epsilon \\
v_i &= \quad -\alpha_i\ (s_{i-1}+x_i) & \\
w_i &= (1+\beta_i)(e_{i-1}+v_i) \quad &; \quad |\beta_i| \le \epsilon \\
s_i &= (1+\gamma_i)(t_i+w_i) \quad &; \quad |\gamma_i| \le \epsilon \\
e_i &= \quad -\gamma_i\ (t_i+w_i) &
\end{aligned}
$$

$$
s_i + e_i = (1 - \alpha_i\beta_i - \beta_i\gamma_{i-1} - \alpha_i\beta_i\gamma_{i-1})(s_{i-1}+e_{i-1}) + (1-\alpha_i\beta_i)x_i
$$

$$
|s_n + e_n - \tilde{s}| \le \frac{(n-1)\sigma}{1-(n-1)\sigma} \sum_{k=1}^{n} |x_k| + \tau \sum_{k=1}^{n} |x_k| + \frac{(n-1)\sigma\tau}{1-(n-1)\sigma} \sum_{k=1}^{n} |x_k|
$$

$$
\sigma := 2\epsilon^2 + \epsilon^3 \quad ; \quad \tau := \epsilon^2
$$

These relations remain valid with subnormal numbers.

What does this buy us?

Table 2: Bounds for $\frac{|s_n+e_n-\tilde{s}|}{\sum_{k=1}^{n}|x_k|}$ using various algorithms at FP64.

	no compensation		6op		double 6op	
n	derived	observed	derived	observed	derived	observed
2^2	4.44E-16	3.08E-17	1.11E-16	0.00E+00	8.63E-32	0.00E+00
2^4	1.78E-15	4.73E-53	1.11E-16	1.32E-69	3.82E-31	1.32E-69
2^6	7.11E-15	4.46E-17	1.11E-16	5.69E-20	1.57E-30	9.36E-34
2^8	2.84E-14	6.97E-17	1.11E-16	3.94E-18	6.30E-30	1.19E-32
2^{10}	1.14E-13	1.57E-16	1.11E-16	1.96E-18	2.52E-29	3.49E-33
2^{12}	4.55E-13	3.73E-16	1.11E-16	5.16E-18	1.01E-28	1.10E-32
2^{14}	1.82E-12	2.19E-17	1.11E-16	1.51E-17	4.04E-28	2.96E-34
2^{16}	7.28E-12	1.78E-17	1.11E-16	9.22E-20	1.62E-27	5.41E-34
2^{18}	2.91E-11	2.66E-17	1.11E-16	2.99E-19	6.46E-27	2.23E-33
2^{20}	1.16E-10	3.58E-17	1.11E-16	1.17E-19	2.58E-26	5.81E-33

Other applications

- Monte Calo simulations (accumulation)
- Dynamical systems
  - Three-body problem (celestial mechanics; chaotic)
  - Periodic orbits in a plane resembling the figure-eight shape
  - ODE system (P_i: position; V_i: velocity; A_i: acceleration)

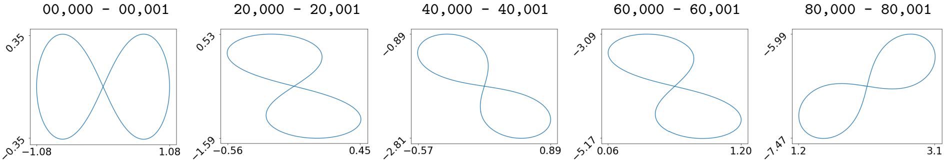
$$\begin{cases} \dot{P}_i &= V_i \\ \dot{V}_i &= A_i = \sum_{j \neq i} \frac{P_j - P_i}{\|P_j - P_i\|^3} \end{cases}$$

  - Time integration

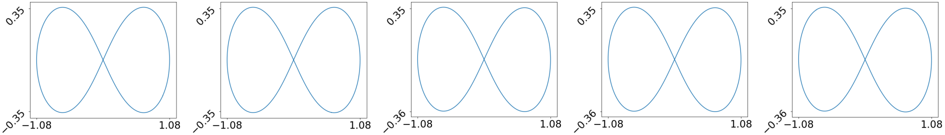
$$\begin{cases} P_i^{(k+1)} &= P_i^{(k)} + hV_i^{(k)} \\ V_i^{(k+1)} &= V_i^{(k)} + hA_i^{(k+1)} \end{cases}$$

  - FP32; $h = 2^{-11}$; 10,000 periods; $\sim$12,955 steps each period

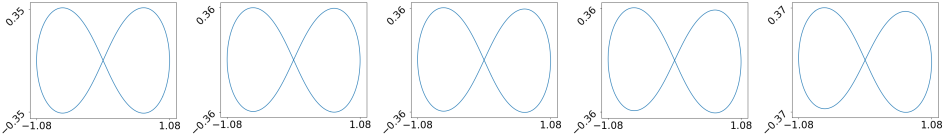
Other applications



Algorithm 1 (no compensation).



Algorithm 2 (6op compensation).



Algorithm 3 (double 6op compensation).

Acknowledgement

This research used resources of the Argonne Leadership Computing Facility, a U.S. Department of Energy (DOE) Office of Science user facility at Argonne National Laboratory and is based on research supported by the U.S. DOE Office of Science-Advanced Scientific Computing Research Program, under Contract No. DE-AC02-06CH11357.

Argonne
NATIONAL LABORATORY

Argonne Leadership Computing Facility