Math 504 Fall 2018 Notes
IEEE Floating Point Arithmetic

Emre Mengi

Department of Mathematics
Koç University
Istanbul, Turkey

IEEE Double Precision Arithmetic

64 binary digits (bits) for each floating point number

$$f = \pm (1.b_1 b_2 \dots b_{52})_2 \times 2^{(a_1 a_2 \dots a_{11})_2}$$

- 52 bits for the significand (mantissa)
- 11 bits for the exponent
- 1 bit for the sign

64 binary digits (bits) for each floating point number

$$f = \pm (1.b_1 b_2 \dots b_{52})_2 \times 2^{(a_1 a_2 \dots a_{11})_2}$$

- 52 bits for the significand (mantissa)
- 11 bits for the exponent
- 1 bit for the sign

Single Precision

32 binary digits for each floating point number

$$f = \pm (1.b_1 b_2 \dots b_{23})_2 \times 2^{(a_1 a_2 \dots a_8)_2}$$

- 23 bits for the significand (mantissa)
- 8 bits for the exponent
- 1 bit for the sign

Single Precision

32 binary digits for each floating point number

$$f = \pm (1.b_1 b_2 \dots b_{23})_2 \times 2^{(a_1 a_2 \dots a_8)_2}$$

- 23 bits for the significand (mantissa)
- 8 bits for the exponent
- 1 bit for the sign

Single Precision

32 binary digits for each floating point number

$$f = \pm (1.b_1 b_2 \dots b_{23})_2 \times 2^{(a_1 a_2 \dots a_8)_2}$$

- 23 bits for the significand (mantissa)
- 8 bits for the exponent
- 1 bit for the sign

The Exponent, $2^{(a_1 a_2 \ldots a_{11})_2}$

- 11 bits can represent $2^{11} = 2048$ exponent values.
- $(00\ldots0)_2$ and $(11\ldots1)_2$ for special purposes
  - $(11\ldots1)_2$ for ∞ and *NaN* (not a number *e.g.* $\infty - \infty$).
  - $(00\ldots0)_2$ for subnormalized representation
- The remaining 2046 exponent values represent any integer in $[-1022, 1023]$.

The Exponent, $2^{(a_1 a_2 \ldots a_{11})_2}$

- 11 bits can represent $2^{11} = 2048$ exponent values.
- $(00\ldots0)_2$ and $(11\ldots1)_2$ for special purposes
  - $(11\ldots1)_2$ for ∞ and *NaN* (not a number *e.g.* $\infty - \infty$).
  - $(00\ldots0)_2$ for subnormalized representation
- The remaining 2046 exponent values represent any integer in $[-1022, 1023]$.

The Exponent, $2^{(a_1 a_2 \ldots a_{11})_2}$

- 11 bits can represent $2^{11} = 2048$ exponent values.
- $(00\ldots0)_2$ and $(11\ldots1)_2$ for special purposes
  - $(11\ldots1)_2$ for ∞ and *NaN* (not a number *e.g.* $\infty - \infty$).
  - $(00\ldots0)_2$ for subnormalized representation
- The remaining 2046 exponent values represent any integer in $[-1022, 1023]$.

The Exponent, $2^{(a_1 a_2 \ldots a_{11})_2}$

- 11 bits can represent $2^{11} = 2048$ exponent values.
- $(00\ldots0)_2$ and $(11\ldots1)_2$ for special purposes
  - $(11\ldots1)_2$ for ∞ and *NaN* (not a number *e.g.* $\infty - \infty$).
  - $(00\ldots0)_2$ for subnormalized representation
- The remaining 2046 exponent values represent any integer in $[-1022, 1023]$.

The Maximal Representable Number

Let x be any floating point number in double precision.

$$\begin{aligned} -(1.11\ldots1)_2 \times 2^{1023} &\leq x \leq (1.11\ldots1)_2 2^{1023} \\ -((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} &\leq x \leq ((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} \\ \underbrace{-(2-2^{-52}) \times 2^{1023}}_{R_{\min}} &\leq x \leq \underbrace{(2-2^{-52}) \times 2^{1023}}_{R_{\max}} \approx 1.8 \times 10^{308} \end{aligned}$$

Thus $x \in \left[-1.8 \cdot 10^{308}, 1.8 \cdot 10^{308}\right]$ roughly.

The Maximal Representable Number

Let x be any floating point number in double precision.

$$\begin{aligned} -(1.11\ldots1)_2 \times 2^{1023} &\leq x \leq (1.11\ldots1)_2 2^{1023} \\ -((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} &\leq x \leq ((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} \\ \underbrace{-(2-2^{-52}) \times 2^{1023}}_{R_{\min}} &\leq x \leq \underbrace{(2-2^{-52}) \times 2^{1023}}_{R_{\max}} \approx 1.8 \times 10^{308} \end{aligned}$$

Thus $x \in \left[-1.8 \cdot 10^{308}, 1.8 \cdot 10^{308}\right]$ roughly.

The Maximal Representable Number

Let x be any floating point number in double precision.

$$-(1.11\ldots1)_2 \times 2^{1023} \leq x \leq (1.11\ldots1)_2 2^{1023}$$

$$-((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} \leq x \leq ((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023}$$

$$\underbrace{-(2-2^{-52}) \times 2^{1023}}_{R_{\min}} \leq x \leq \underbrace{(2-2^{-52}) \times 2^{1023}}_{R_{\max}} \approx 1.8 \times 10^{308}$$

Thus $x \in \left[-1.8 \cdot 10^{308}, 1.8 \cdot 10^{308}\right]$ roughly.

The Maximal Representable Number

Let x be any floating point number in double precision.

$$\begin{aligned} -(1.11\ldots1)_2 \times 2^{1023} &\leq x \leq (1.11\ldots1)_2 2^{1023} \\ -((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} &\leq x \leq ((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} \\ \underbrace{-(2-2^{-52}) \times 2^{1023}}_{R_{\min}} &\leq x \leq \underbrace{(2-2^{-52}) \times 2^{1023}}_{R_{\max}} \approx 1.8 \times 10^{308} \end{aligned}$$

Thus $x \in \left[-1.8 \cdot 10^{308}, 1.8 \cdot 10^{308}\right]$ roughly.

The Maximal Representable Number

Let x be any floating point number in double precision.

$$\begin{array}{rcl} -(1.11\ldots1)_2 \times 2^{1023} & \leq x \leq & (1.11\ldots1)_2 2^{1023} \\ -((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} & \leq x \leq & ((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} \\ \underbrace{-(2-2^{-52}) \times 2^{1023}}_{R_{\min}} & \leq x \leq & \underbrace{(2-2^{-52}) \times 2^{1023}}_{R_{\max}} \approx 1.8 \times 10^{308} \end{array}$$

Thus $x \in [-1.8 \cdot 10^{308}, 1.8 \cdot 10^{308}]$ roughly.

The Maximal Representable Number

Let x be any floating point number in double precision.

$$\begin{array}{rcl} -(1.11\ldots1)_2 \times 2^{1023} & \leq x \leq & (1.11\ldots1)_2 2^{1023} \\ -((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} & \leq x \leq & ((10.0\ldots0)_2 - (0.0\ldots1)_2) \times 2^{1023} \\ \underbrace{-(2-2^{-52}) \times 2^{1023}}_{R_{\min}} & \leq x \leq & \underbrace{(2-2^{-52}) \times 2^{1023}}_{R_{\max}} \approx 1.8 \times 10^{308} \end{array}$$

Thus $x \in \left[-1.8 \cdot 10^{308}, 1.8 \cdot 10^{308}\right]$ roughly.

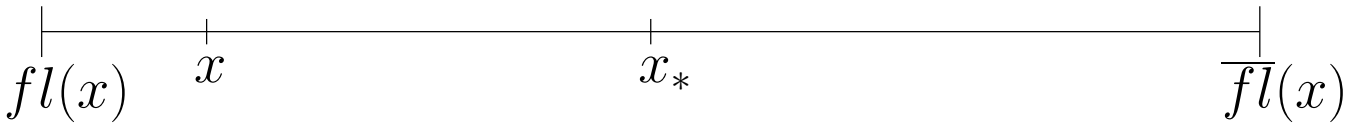
The Machine Precision

Let $x = s \cdot 2^E \in [R_{\min}, R_{\max}]$

Let $fl(x) = \hat{s} \cdot 2^E$ be the floating point number closest to x

Relative error — $E(x) := \frac{|x - fl(x)|}{|x|}$

The Machine Precision

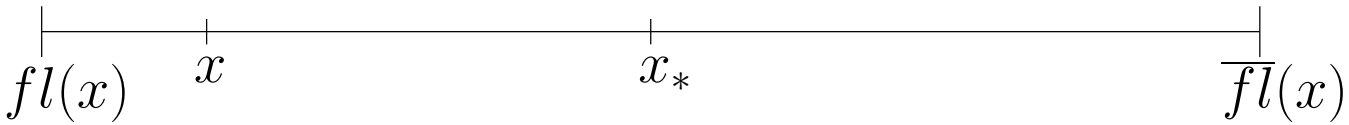


Let $x = s \cdot 2^E \in [R_{\min}, R_{\max}]$

Let $fl(x) = \hat{s} \cdot 2^E$ be the floating point number closest to x

Relative error — $E(x) := \frac{|x - fl(x)|}{|x|}$

The Machine Precision

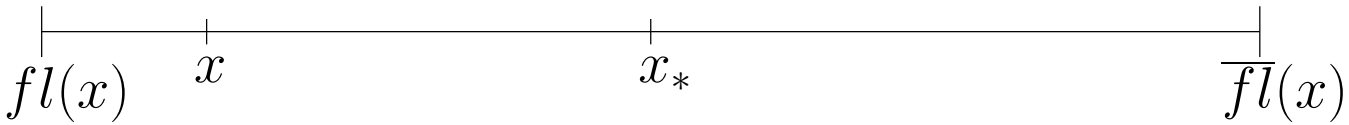


Let $x = s \cdot 2^E \in [R_{\min}, R_{\max}]$

Let $fl(x) = \hat{s} \cdot 2^E$ be the floating point number closest to x

Relative error — $E(x) := \frac{|x - fl(x)|}{|x|}$

The Machine Precision

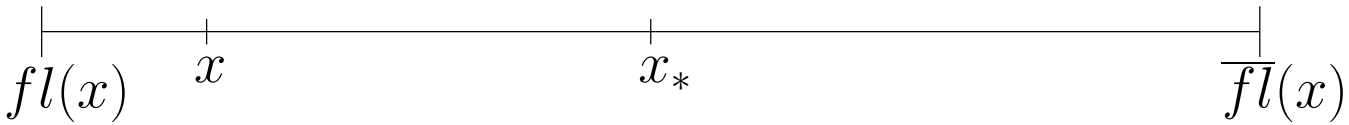


Let $x = s \cdot 2^E \in [R_{\min}, R_{\max}]$

Let $fl(x) = \hat{s} \cdot 2^E$ be the floating point number closest to x

Relative error — $E(x) := \frac{|x - fl(x)|}{|x|}$

The Machine Precision

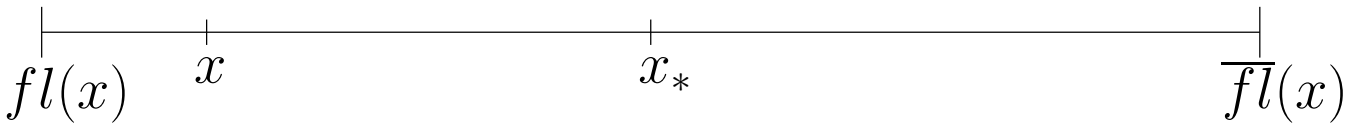


Let $x = s \cdot 2^E \in [R_{\min}, R_{\max}]$

Let $fl(x) = \hat{s} \cdot 2^E$ be the floating point number closest to x

Relative error — $E(x) := \frac{|x - fl(x)|}{|x|}$

The Machine Precision



Let $x = s \cdot 2^E \in [R_{\min}, R_{\max}]$

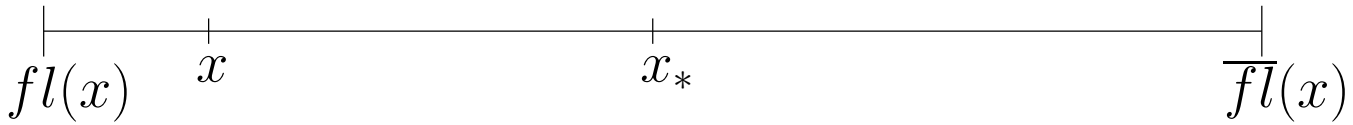
Let $fl(x) = \hat{s} \cdot 2^E$ be the floating point number closest to x

Relative error — $E(x) := \frac{|x - fl(x)|}{|x|}$

ϵ_{mach} (machine precision)

Maximal relative error — $\epsilon_{\text{mach}} := \max_{x \in [R_{\min}, R_{\max}]} E(x)$

The Machine Precision



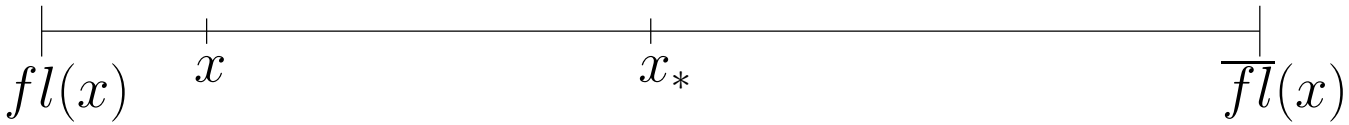
$\overline{fl}(x) = (\hat{s} + 2^{-52}) \cdot 2^E$ and $x_* = \frac{fl(x)+\overline{fl}(x)}{2} = (\hat{s} + 2^{-53}) \cdot 2^E$

Bounding the relative error

$$E(x) = \frac{|x - fl(x)|}{|x|} \leq \frac{|x_* - fl(x)|}{|fl(x)|} = \frac{2^{-53} \cdot 2^E}{\hat{s} \cdot 2^E} \leq 2^{-53}$$

Thus $\epsilon_{\text{mach}} = 2^{-53} \approx 1.11 \cdot 10^{-16}$.

The Machine Precision



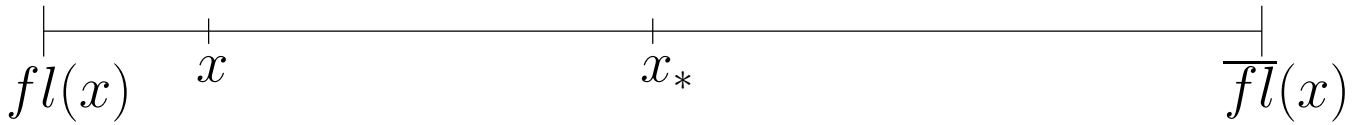
$\overline{fl}(x) = (\hat{s} + 2^{-52}) \cdot 2^E$ and $x_* = \frac{fl(x)+\overline{fl}(x)}{2} = (\hat{s} + 2^{-53}) \cdot 2^E$

Bounding the relative error

$$E(x) = \frac{|x - fl(x)|}{|x|} \leq \frac{|x_* - fl(x)|}{|fl(x)|} = \frac{2^{-53} \cdot 2^E}{\hat{s} \cdot 2^E} \leq 2^{-53}$$

Thus $\epsilon_{\text{mach}} = 2^{-53} \approx 1.11 \cdot 10^{-16}$.

The Machine Precision



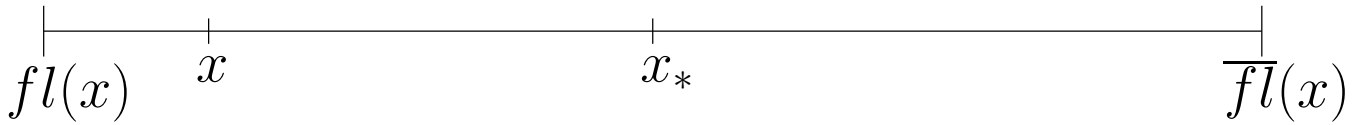
$\overline{fl}(x) = (\hat{s} + 2^{-52}) \cdot 2^E$ and $x_* = \frac{fl(x)+\overline{fl}(x)}{2} = (\hat{s} + 2^{-53}) \cdot 2^E$

Bounding the relative error

$$E(x) = \frac{|x - fl(x)|}{|x|} \leq \frac{|x_* - fl(x)|}{|fl(x)|} = \frac{2^{-53} \cdot 2^E}{\hat{s} \cdot 2^E} \leq 2^{-53}$$

Thus $\epsilon_{\text{mach}} = 2^{-53} \approx 1.11 \cdot 10^{-16}$.

The Machine Precision



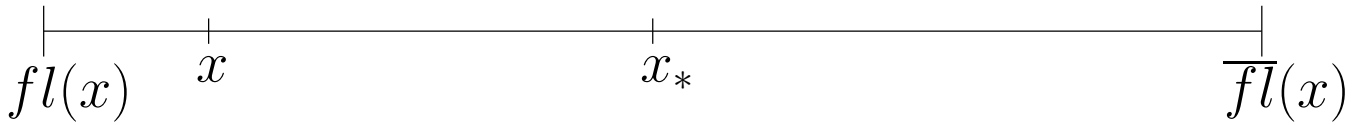
$\overline{fl}(x) = (\hat{s} + 2^{-52}) \cdot 2^E$ and $x_* = \frac{fl(x)+\overline{fl}(x)}{2} = (\hat{s} + 2^{-53}) \cdot 2^E$

Bounding the relative error

$$E(x) = \frac{|x - fl(x)|}{|x|} \leq \frac{|x_* - fl(x)|}{|fl(x)|} = \frac{2^{-53} \cdot 2^E}{\hat{s} \cdot 2^E} \leq 2^{-53}$$

Thus $\epsilon_{\text{mach}} = 2^{-53} \approx 1.11 \cdot 10^{-16}$.

The Machine Precision



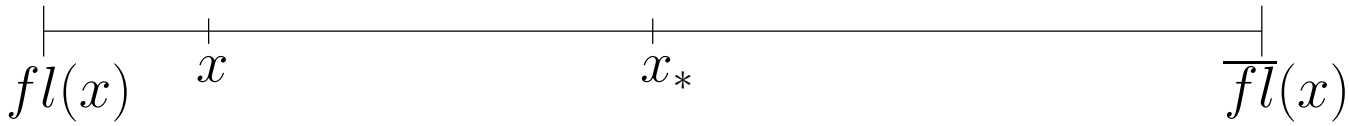
$\overline{fl}(x) = (\hat{s} + 2^{-52}) \cdot 2^E$ and $x_* = \frac{fl(x)+\overline{fl}(x)}{2} = (\hat{s} + 2^{-53}) \cdot 2^E$

Bounding the relative error

$$E(x) = \frac{|x - fl(x)|}{|x|} \leq \frac{|x_* - fl(x)|}{|fl(x)|} = \frac{2^{-53} \cdot 2^E}{\hat{s} \cdot 2^E} \leq 2^{-53}$$

Thus $\epsilon_{\text{mach}} = 2^{-53} \approx 1.11 \cdot 10^{-16}$.

The Machine Precision



$\overline{fl}(x) = (\hat{s} + 2^{-52}) \cdot 2^E$ and $x_* = \frac{fl(x)+\overline{fl}(x)}{2} = (\hat{s} + 2^{-53}) \cdot 2^E$

Bounding the relative error

$$E(x) = \frac{|x - fl(x)|}{|x|} \leq \frac{|x_* - fl(x)|}{|fl(x)|} = \frac{2^{-53} \cdot 2^E}{\hat{s} \cdot 2^E} \leq 2^{-53}$$

Thus $\epsilon_{\text{mach}} = 2^{-53} \approx 1.11 \cdot 10^{-16}$.

Performing Flops in IEEE Double Precision

The Floating Point Operations (flops)

The floating point operations or flops $(\oplus, \otimes, \ominus, \oslash)$ must satisfy the following for each $x, y \in [R_{\min}, R_{\max}]$:

$$x \oplus y = fl(x + y)$$
$$x \ominus y = fl(x - y)$$
$$x \otimes y = fl(x \times y)$$
$$x \oslash y = fl(x / y)$$

The Floating Point Operations (flops)

The floating point operations or flops $(\oplus, \otimes, \ominus, \oslash)$ must satisfy the following for each $x, y \in [R_{\min}, R_{\max}]$:

$$x \oplus y = fl(x + y)$$
$$x \ominus y = fl(x - y)$$
$$x \otimes y = fl(x \times y)$$
$$x \oslash y = fl(x / y)$$

The Floating Point Operations (flops)

The floating point operations or flops $(\oplus, \otimes, \ominus, \oslash)$ must satisfy the following for each $x, y \in [R_{\min}, R_{\max}]$:

$$
\begin{aligned}
x \oplus y &= fl(x + y) \\
x \ominus y &= fl(x - y) \\
x \otimes y &= fl(x \times y) \\
x \oslash y &= fl(x / y)
\end{aligned}
$$

The Floating Point Operations (flops)

Ex 1

$$1 \oplus 2^{-52} = 1 + 2^{-52}, \qquad \text{but } 1 \oplus 2^{-54} = 1$$

Ex 2

$$\begin{aligned} (1 + 2^{-52}) \otimes (2 + 2^{-51}) &= fl(2 + 2^{-51} + 2^{-51} + 2^{-103}) \\ &= fl\left\{2(1 + 2^{-52} + 2^{-52} + 2^{-104})\right\} \\ &= 2(1 + 2^{-51}) \end{aligned}$$

The Floating Point Operations (flops)

Ex 1

$$1 \oplus 2^{-52} = 1 + 2^{-52}, \qquad \text{but } 1 \oplus 2^{-54} = 1$$

Ex 2

$$\begin{aligned} (1 + 2^{-52}) \otimes (2 + 2^{-51}) &= fl(2 + 2^{-51} + 2^{-51} + 2^{-103}) \\ &= fl\left\{2(1 + 2^{-52} + 2^{-52} + 2^{-104})\right\} \\ &= 2(1 + 2^{-51}) \end{aligned}$$

The Floating Point Operations (flops)

Ex 1

$$1 \oplus 2^{-52} = 1 + 2^{-52}, \qquad \text{but } 1 \oplus 2^{-54} = 1$$

Ex 2

$$\begin{aligned} (1 + 2^{-52}) \otimes (2 + 2^{-51}) &= fl(2 + 2^{-51} + 2^{-51} + 2^{-103}) \\ &= fl\left\{2(1 + 2^{-52} + 2^{-52} + 2^{-104})\right\} \\ &= 2(1 + 2^{-51}) \end{aligned}$$

The Floating Point Operations (flops)

Ex 1

$$1 \oplus 2^{-52} = 1 + 2^{-52}, \qquad \text{but } 1 \oplus 2^{-54} = 1$$

Ex 2

$$\begin{aligned}(1 + 2^{-52}) \otimes (2 + 2^{-51}) &= fl(2 + 2^{-51} + 2^{-51} + 2^{-103}) \\ &= fl\left\{2(1 + 2^{-52} + 2^{-52} + 2^{-104})\right\} \\ &= 2(1 + 2^{-51})\end{aligned}$$

The Floating Point Operations (flops)

Ex 1

$$1 \oplus 2^{-52} = 1 + 2^{-52}, \qquad \text{but } 1 \oplus 2^{-54} = 1$$

Ex 2

$$\begin{aligned} (1 + 2^{-52}) \otimes (2 + 2^{-51}) &= fl(2 + 2^{-51} + 2^{-51} + 2^{-103}) \\ &= fl\left\{2(1 + 2^{-52} + 2^{-52} + 2^{-104})\right\} \\ &= 2(1 + 2^{-51}) \end{aligned}$$

The Floating Point Operations (flops)

Ex 1

$$1 \oplus 2^{-52} = 1 + 2^{-52}, \qquad \text{but } 1 \oplus 2^{-54} = 1$$

Ex 2

$$\begin{aligned} (1 + 2^{-52}) \otimes (2 + 2^{-51}) &= fl(2 + 2^{-51} + 2^{-51} + 2^{-103}) \\ &= fl\left\{2(1 + 2^{-52} + 2^{-52} + 2^{-104})\right\} \\ &= 2(1 + 2^{-51}) \end{aligned}$$

The Floating Point Operations (flops)

Ex 1

$$1 \oplus 2^{-52} = 1 + 2^{-52}, \qquad \text{but } 1 \oplus 2^{-54} = 1$$

Ex 2

$$\begin{aligned} (1 + 2^{-52}) \otimes (2 + 2^{-51}) &= fl(2 + 2^{-51} + 2^{-51} + 2^{-103}) \\ &= fl\left\{2(1 + 2^{-52} + 2^{-52} + 2^{-104})\right\} \\ &= 2(1 + 2^{-51}) \end{aligned}$$