

# Exercise 1: Floating-Point Arithmetic

## IEEE 754 Double Precision Analysis

### 1 Main Problem (4 points)

#### Problem Statement

Let  $x$  be a floating-point number in IEEE 754 double precision such that  $1 \leq x < 2$ . Show that, when rounding to the nearest,  $\text{fl}(x \times \text{fl}(1/x))$  is either 1 or  $1 - u$ , where  $u = 2^{-53}$ .

#### Solution

##### Setup

- $x \in \mathbb{F}$  (floating-point number in IEEE 754 double precision)
- $1 \leq x < 2$
- Rounding to nearest
- Unit roundoff:  $u = 2^{-53}$

**Goal:** Show that  $\text{fl}(x \times \text{fl}(1/x)) \in \{1, 1 - u\}$

##### Key Observations

**Representable numbers in  $[1, 2)$**  In the interval  $[1, 2)$ , floating-point numbers have the form:

$$x = 1.f \times 2^0 = 1 + f \times 2^{-52} \quad (1)$$

where  $f \in \{0, 1, 2, \dots, 2^{52} - 1\}$  (52-bit fraction).

The **spacing** (gap between consecutive numbers) is  $2^{-52}$ .

**Unit in Last Place (ULP)** For  $x \in [1, 2)$ :

- $\text{ulp}(x) = 2^{-52}$
- $u = 2^{-53} = \frac{1}{2}\text{ulp}(x)$

##### Step 1: Analyze $\text{fl}(1/x)$

Since  $1 \leq x < 2$ , we have:

$$\frac{1}{2} < \frac{1}{x} \leq 1 \quad (2)$$

Using the standard model with rounding to nearest:

$$\text{fl}(1/x) = \frac{1}{x}(1 + \delta_1) \quad \text{where } |\delta_1| \leq u \quad (3)$$

**Step 2: Analyze  $\text{fl}(x \times \text{fl}(1/x))$** 

Let  $y = \text{fl}(1/x)$ . Then:

$$\text{fl}(x \times y) = (x \times y)(1 + \delta_2) \quad \text{where } |\delta_2| \leq u \quad (4)$$

Substituting  $y = \frac{1}{x}(1 + \delta_1)$ :

$$\text{fl}(x \times y) = x \cdot \frac{1}{x}(1 + \delta_1) \cdot (1 + \delta_2) \quad (5)$$

$$= (1 + \delta_1)(1 + \delta_2) \quad (6)$$

$$= 1 + \delta_1 + \delta_2 + \delta_1\delta_2 \quad (7)$$

**Step 3: Bound the Result**

Since  $|\delta_1|, |\delta_2| \leq u = 2^{-53}$ :

$$|(\delta_1 + \delta_2 + \delta_1\delta_2)| \leq |\delta_1| + |\delta_2| + |\delta_1\delta_2| \quad (8)$$

$$\leq u + u + u^2 = 2u + u^2 \quad (9)$$

Since  $u^2 = 2^{-106}$  is negligible compared to  $u = 2^{-53}$ :

$$|(\delta_1 + \delta_2 + \delta_1\delta_2)| \leq 2u + u^2 < 3u \quad (10)$$

**Step 4: Determine Possible Values**

The exact value is:

$$\text{result} = 1 + \varepsilon \quad \text{where } |\varepsilon| < 3u \quad (11)$$

The representable numbers near 1:

- In  $[1/2, 1)$ : spacing is  $2^{-53} = u$
- At exactly 1: this is the boundary
- In  $[1, 2)$ : spacing is  $2^{-52} = 2u$

So near 1:

- $1 - u$  (last number before 1 in the interval  $[1/2, 1)$ )
- 1 (exact)
- $1 + 2u$  (first number after 1 in the interval  $[1, 2)$ )

**Step 5: Apply Rounding to Nearest**

We have  $\text{result} = 1 + \varepsilon$  where  $|\varepsilon| < 3u < 1.5 \times 2u$ .

Given that:

$$\text{fl}(x \times \text{fl}(1/x)) = 1 + \varepsilon \quad \text{where } |\varepsilon| < 3u \quad (12)$$

and the representable numbers near 1 are  $1 - u$ , 1, and  $1 + 2u$ , the result after rounding to nearest is:

Answer

$$\boxed{\mathfrak{fl}(x \times \mathfrak{fl}(1/x)) \in \{1 - u, 1\}} \quad (13)$$

where  $u = 2^{-53}$ .

## 2 Variation 1: Directed Rounding

### Problem Statement

Let  $x$  be a floating-point number in IEEE 754 double precision such that  $1 \leq x < 2$ . When using rounding toward  $+\infty$  (instead of rounding to nearest), show that  $\text{fl}_\uparrow(x \times \text{fl}_\uparrow(1/x)) \geq 1$  and bound the maximum value it can achieve.

### Key Difference

- Both operations round upward
- $\text{fl}_\uparrow(1/x) \geq 1/x$ , so the product tends to be  $\geq 1$
- Need to find the worst-case upper bound

### Solution

#### Step 1: Compute $\text{fl}_\uparrow(1/x)$

With directed rounding upward:

$$\text{fl}_\uparrow(1/x) = \frac{1}{x}(1 + \delta_1), \quad 0 \leq \delta_1 \leq u \quad (14)$$

**Note:**  $\delta_1 \geq 0$  because we always round up.

#### Step 2: Compute $\text{fl}_\uparrow(x \times \text{fl}_\uparrow(1/x))$

$$\text{fl}_\uparrow(x \times \text{fl}_\uparrow(1/x)) = x \cdot \frac{1}{x}(1 + \delta_1) \cdot (1 + \delta_2) \quad (15)$$

where  $0 \leq \delta_2 \leq u$  (rounding up again).

$$= (1 + \delta_1)(1 + \delta_2) \quad (16)$$

#### Step 3: Bound the Result

**Minimum value** (when  $\delta_1 = \delta_2 = 0$ ):

$$\text{result}_{\min} = 1 \quad (17)$$

**Maximum value** (when  $\delta_1 = \delta_2 = u$ ):

$$\text{result}_{\max} = (1 + u)^2 = 1 + 2u + u^2 \quad (18)$$

$$\approx 1 + 2u \quad (19)$$

Since  $u^2 = 2^{-106} \ll u = 2^{-53}$ , we have:

$$\text{result}_{\max} < 1 + 2u + u = 1 + 3u \quad (20)$$

#### Step 4: Identify Possible Values

After rounding up to nearest representable number in  $[1, 2)$  with spacing  $2u$ :

- 1 (exact, if result is exactly 1)
- $1 + 2u$  (next representable number)
- Possibly  $1 + 4u$  if result  $> 1 + 2u$

Since result  $\leq 1 + 2u + u^2 < 1 + 2u + u$ , and this must round to a representable value:

Answer

$$1 \leq \text{fl}_\uparrow(x \times \text{fl}_\uparrow(1/x)) \leq 1 + 2u \quad (21)$$

Possible values:  $\{1, 1 + 2u\}$

### 3 Variation 2: Multiple Operations

#### Problem Statement

Let  $x$  be a floating-point number in IEEE 754 double precision such that  $1 \leq x < 2$ . Show that when rounding to the nearest:

$$\text{fl}\left(\frac{x}{\text{fl}(x+1)}\right) = \text{fl}(x) \times \text{fl}\left(\frac{1}{\text{fl}(x+1)}\right) + \delta \quad (22)$$

where  $|\delta| \leq Cu$  for some small constant  $C$ , and determine the set of possible values.

#### Key Difference

- Involves addition, division, and multiplication
- Tests understanding of error propagation through multiple operations
- Requires analyzing  $x/(x+1)$  which is in  $[1/2, 2/3)$  for  $x \in [1, 2)$

#### Solution

**Step 1: Compute  $s = \text{fl}(x+1)$**

Since  $x \in [1, 2)$ :

$$s = \text{fl}(x+1) = (x+1)(1+\delta_1), \quad |\delta_1| \leq u \quad (23)$$

So  $s \in [2, 3)$  approximately.

**Step 2: Compute  $\text{fl}(1/s)$**

$$\text{fl}(1/s) = \frac{1}{s}(1+\delta_2) = \frac{1}{(x+1)(1+\delta_1)}(1+\delta_2) \quad (24)$$

$$= \frac{1}{x+1} \cdot \frac{1+\delta_2}{1+\delta_1} \quad (25)$$

Using  $\frac{1}{1+\delta_1} \approx 1 - \delta_1$  for small  $\delta_1$ :

$$\text{fl}(1/s) \approx \frac{1}{x+1}(1+\delta_2 - \delta_1 - \delta_1\delta_2) \quad (26)$$

$$\approx \frac{1}{x+1}(1+\delta_2 - \delta_1) \quad (27)$$

**Step 3: Compute  $\text{fl}(x \times \text{fl}(1/s))$**

$$\text{fl}(x \times \text{fl}(1/s)) = x \cdot \frac{1}{x+1}(1+\delta_2 - \delta_1) \cdot (1+\delta_3) \quad (28)$$

where  $|\delta_3| \leq u$ .

$$= \frac{x}{x+1}(1+\delta_2 - \delta_1)(1+\delta_3) \quad (29)$$

$$= \frac{x}{x+1}(1+\delta_2 - \delta_1 + \delta_3 + \text{higher order terms}) \quad (30)$$

**Step 4: Bound the Total Error**

The exact value is  $\frac{x}{x+1}$ , and the computed value has relative error:

$$\varepsilon_{\text{total}} = \delta_2 - \delta_1 + \delta_3 + O(u^2) \quad (31)$$

$$|\varepsilon_{\text{total}}| \leq 3u + O(u^2) \quad (32)$$

**Step 5: Determine Range**

For  $x \in [1, 2)$ :

$$\frac{x}{x+1} \in \left[ \frac{1}{2}, \frac{2}{3} \right) \quad (33)$$

The computed result satisfies:

$$\text{result} = \frac{x}{x+1}(1 + \varepsilon) \quad \text{where } |\varepsilon| \leq 3u \quad (34)$$

**Answer**

The result is within relative error  $3u$  of  $\frac{x}{x+1}$ . The set of possible values depends on the specific representable numbers in  $[1/2, 2/3)$ , but the error bound is:

$$\left| \text{fl} \left( x \times \text{fl} \left( \frac{1}{\text{fl}(x+1)} \right) \right) - \frac{x}{x+1} \right| \leq 3u \cdot \frac{x}{x+1} + O(u^2) \quad (35)$$

## 4 Variation 3: Fused Multiply-Add

### Problem Statement

Let  $x$  be a floating-point number in IEEE 754 double precision such that  $1 \leq x < 2$ . Compare:

1.  $\text{fl}(x \times \text{fl}(1/x))$  (standard computation)
2.  $\text{FMA}(x, \text{fl}(1/x), 0)$  (using fused multiply-add)

Show that with FMA, the result is **always exactly** 1.

### Key Difference

- FMA performs  $x \times y + z$  with only **one rounding** at the end
- Eliminates one source of error

### Solution

#### Standard Computation

$$\text{fl}(x \times \text{fl}(1/x)) = (1 + \delta_1)(1 + \delta_2) = 1 + \delta_1 + \delta_2 + \delta_1\delta_2 \quad (36)$$

**Result:** Can be  $1 - u$  or 1 (as shown in original problem).

#### With FMA

$$\text{FMA}(x, \text{fl}(1/x), 0) = \text{fl}(x \times \text{fl}(1/x)) \quad (37)$$

But now with only **one rounding**:

$$= \text{fl}\left(x \times \frac{1}{x}(1 + \delta_1)\right) = \text{fl}(1 + \delta_1) \quad (38)$$

where  $|\delta_1| \leq u$ .

Since  $|1 + \delta_1 - 1| = |\delta_1| \leq u$ , and the nearest representable values are  $1 - u$ , 1, and  $1 + 2u$ :

- If  $|\delta_1| \leq u$ , the value  $1 + \delta_1$  is within  $u$  of 1
- Rounding to nearest gives exactly **1**

### Answer

With FMA:

$$\boxed{\text{FMA}(x, \text{fl}(1/x), 0) = 1 \text{ always (exact)}} \quad (39)$$

This demonstrates the accuracy improvement from FMA: eliminates one rounding operation, reducing error from  $O(2u)$  to exact.

Method	Possible Values	Error Bound
Round to nearest	$\{1 - u, 1\}$	$ \varepsilon  < 3u$
Round upward	$\{1, 1 + 2u\}$	$0 \leq \varepsilon \leq 2u + u^2$
With addition $x/(x + 1)$	Multiple values	$ \varepsilon  \leq 3u$ (relative)
With FMA	$\{1\}$ (exact)	0 (exact)

Table 1: Comparison of different computation methods

## 5 Summary Comparison

### Key Insights

1. **Standard rounding:** Two rounding operations introduce  $O(2u)$  error
2. **Directed rounding:** Bias ensures result  $\geq 1$ , but can be larger
3. **Multiple operations:** Error accumulates linearly with number of operations
4. **FMA advantage:** Single rounding eliminates compound errors, achieving exact results