When (3x/3) and 3(x/3) are not equal to x

Frédéric Goualard

Team OGRE

When $(3x)/3$ and $3(x/3)$ are not equal to x

Rapport de recherche
Research Report

RR n° xx.xx

Frédéric Goualard

LS2N
LABORATOIRE DES SCIENCES DU NUMÉRIQUE DE NANTES

Frédéric Goualard
When $(3x)/3$ *and* $3(x/3)$ *are not equal to* x

Les rapports de recherche du Laboratoire des Sciences du Numérique de Nantes sont disponibles au format Adobe® PDF sur http://ls2n.fr/.

Research Reports from the Laboratoire des Sciences du Numérique de Nantes are available at http://ls2n.fr/.

When $(3x)/3$ and $3(x/3)$ are not equal to x

Frédéric Goualard

Abstract

Rounding Error Analysis is routinely used to compute a worst-case error bound on the result of algorithms that use floating-point arithmetic. However, for some applications (e.g., when it is necessary to prove some inclusion of the result in a domain), the knowledge of both an upper-bound of the magnitude of the error and of its sign is paramount. Using standard rounding error analysis together with a simple systematic approach, we compute such information for the expressions $3x$, $x/3$, $3(x/3)$, and $3x/3$, which can be used, e.g., in the proof of interval arithmetic operators.

CCS Concepts: **Mathematics of computing→Interval arithmetic**, *Theory of computation→Rounding techniques*

Additional Key Words and Phrases: floating-point arithmetic, error analysis, roundoff, interval arithmetic

Contents

1 Introduction

The implementation of Interval Arithmetic operators requires the evaluation of floating-point expressions, which can be marred by rounding errors that jeopardize the very properties that Interval Arithmetic is supposed to ensure [4]. *Rounding error analysis* [6] can be used to determine in a systematic way the magnitude of these errors. It is usually, however, not concerned with their sign, a fatal flaw whenever the question of the inclusion of a value into an interval is to be ascertained.

Some clever use of the binary floating-point number format properties as guaranteed by the IEEE 754 standard [7] allows in special cases to work out the conditions under which rounding errors cancel out in such a way that the computed result is exactly equal to the true result. Several examples of such an approach can be found in Goldberg's survey [3] and Muller *et al.*'s book [9], to name a few sources. On the other hand, few works concern themselves with determining both the magnitude *and* the sign of rounding errors in computing arithmetic expressions.

In this paper, we investigate the case of the expressions $3x/3$ and $3(x/3)$ for x a binary floating-point number. Using a mixture of classical Rounding Analysis and systematic study of the elementary operations "$3x$" and "$x/3$", we determinate for each possible value of x the sign and the magnitude of the error in computing $3x/3$ and $3(x/3)$. This study is used in a forthcoming article [5] to prove inclusion properties of some interval trisection operators. We expect it to be useful in the study of the many other algorithms that rely on a multiplication or a division by 3 (see, e.g., de Dinechin et al.'s work [1]).

2 Binary floating-point numbers

In this study, we only concern ourselves with binary floating-point numbers as defined by the 1985 IEEE 754 standard [7], by far the most widespread standard in use today for floating-point numbers. This paper assumes a prior knowledge of the basics of the IEEE 754 standard even though the most important points are summarized hereafter as needed for the sake of completeness.

We consider binary floating-point numbers from a set $\mathbb{F}$ represented with a significand of size p bits:

$$x \in \mathbb{F}, \quad x = \pm b_0.b_1 \dots b_{p-1} \times 2^e$$

with an exponent e in the range $[E_{\min}, E_{\max}]$.

For the purpose of the systematic study of the elementary operations, we assume $p \geqslant 5$, not a strong constraint considering that the smallest format presented in the IEEE 754 standard has $p = 24$. We will not take into account the possibility of overflow when performing the multiplication by 3; on the other hand, the possibility of underflow will be fully adressed, albeit separately from the normal cases.

When the base b used to represent a number x is not unambiguously drawn from the context, we will indicate it as such: x_b.

All results are supposed to be rounded to nearest-even, the usual default rounding strategy. Given r a real number, we note $\mathrm{fl}(r)$ the floating-point value corresponding to r rounded to nearest-even. Given x, a floating-point number, we note x^- the greatest floating-point number smaller than x, and x^+ the smallest floating-point number greater than x. For an expression of the form $(x \diamond y) \square z$—with $(x, y, z) \in \mathbb{F}^3$ and $(\diamond, \square)$ arithmetic operators—, we use the shorthand $\mathrm{fl}\langle(x \diamond y) \square z\rangle$[1] to mean $\mathrm{fl}\big(\mathrm{fl}\big(x \diamond y\big) \square z\big)$. Lastly, we will extensively use in this paper the fact that rounding is order-preserving (*rounding monotonicity* [6, p. 38]):

$$\forall (r_1, r_2) \in \mathbb{R}^2 : \; r_1 \geqslant r_2 \implies \mathrm{fl}(r_1) \geqslant \mathrm{fl}(r_2) \tag{1}$$

Note that if there is a strict inequality between r_1 and r_2 (i.e., $r_1 > r_2$), we still have $\mathrm{fl}(r_1) \geqslant \mathrm{fl}(r_2)$.

With *correctly rounded* floating-point operators, as specified by the IEEE 754 standard, there is a simple relationship between a computed value and the real result [6]:

$$\mathrm{fl}\big(x \diamond y\big) = (x \diamond y)(1 + \delta) + \eta, \quad \begin{cases} (x, y) \in \mathbb{F}^2, \\ \delta\eta = 0, \\ |\delta| \leqslant u, \\ |\eta| \leqslant \mu/2 \\ \diamond \in \{+, -, \times, \div\} \end{cases} \tag{2}$$

[1] Note the parentheses replaced by chevrons.

where u is the *unit roundoff* equal to 2^{-p} and μ is the *smallest positive subnormal number* equal to $u2^{E_{\min}+1}$. The bound on δ may be somewhat refined depending on the operator; in particular, we will make use of the following improved bounds [8]:

$$\begin{cases} \mathrm{fl}(xy) & = xy(1+\delta) \quad , \quad |\delta| \leqslant \dfrac{u}{1+u} \\ \mathrm{fl}\left(\dfrac{x}{y}\right) & = \dfrac{x}{y}(1+\delta) \quad , \quad |\delta| \leqslant u - 2u^2 \end{cases} \quad , \quad (x,y) \in \mathbb{F}^2, \tag{3}$$

provided no underflow or overflow occurs.

3 Multiplying x by three

Barring overflow, the multiplication of two binary floating-point numbers:

$$x = (-1)^{s_x} m_x \times 2^{e_x}$$

and

$$y = (-1)^{s_y} m_y \times 2^{e_y}$$

is given by [9]:

$$xy = (-1)^{s_x \oplus s_y} m_x m_y \times 2^{e_x + e_y} \tag{4}$$

The multiplication of a floating-point number x by 3 (i.e., 1.1×2^1 in normal floating-point binary form) is particularly easy to perform thanks to the simplicity of the representation of 3. We will consider first the case for which no underflow occurs; denormal numbers will be considered next.

3.1 The normal case

In this section, x is a normal number of the form $x = (-1)^{s_x} 1.b_1 \dots b_{p-1} \times 2^{e_x}$. From Eq. (4), we see that we may consider $x > 0$ only, as the results thus obtained need only be mirrored to obtain their counterpart when $x < 0$.

Depending on x, there are two possibilities: either $3x$ has a $p+1$ bits significand (Eq. (5a)), or it has a $p+2$ bits significand (Eq. (5b)).

$$\begin{array}{cccccccccc} & & 1. & b_1 & b_2 & \cdots & b_{p-3} & b_{p-2} & b_{p-1} & \times 2^{e_x} \\ \times & & 1. & 1 & & & & & & \times 2^{1} \\ \hline & & 1. & b_1 & b_2 & \cdots & b_{p-3} & b_{p-2} & b_{p-1} & \\ + & 1. & b_1 & b_2 & \cdots & b_{p-3} & b_{p-2} & b_{p-1} & & \\ \hline & 1. & b'_1 & b'_2 & \cdots & b'_{p-3} & b'_{p-2} & b'_{p-1} & b'_p & \times 2^{e_x+1} \end{array} \tag{5a}$$

$$\begin{array}{cccccccccc} & & 1. & b_1 & b_2 & \cdots & b_{p-3} & b_{p-2} & b_{p-1} & \times 2^{e_x} \\ \times & & 1. & 1 & & & & & & \times 2^{1} \\ \hline & & 1. & b_1 & b_2 & \cdots & b_{p-3} & b_{p-2} & b_{p-1} & \\ + & 1. & b_1 & b_2 & \cdots & b_{p-3} & b_{p-2} & b_{p-1} & & \\ \hline 1. & b'_1 & b'_2 & \cdots & b'_{p-3} & b'_{p-2} & b'_{p-1} & b'_p & b'_{p+1} & \times 2^{e_x+2} \end{array} \tag{5b}$$

From Eqs. (5a) and (5b), we can deduce conditions on x for $y = 3x$ to have a $p+1$ bits significand; similarly, we can deduce the value of b'_1 depending on the number of bits of $3x$:

Lemma 1. *Given $x = 1.b_1 \dots b_{p-1} \times 2^{e_x}$ a normal binary floating-point number. If $y = 3x$ has a $p+1$ bits significand, then $x = 1.\mathbf{0}b_2 \dots b_{p-1} \times 2^{e_x}$ and $y = 1.\mathbf{1}b'_2 \dots b'_{p-1}b'_p \times 2^{e_x+1}$*

Proof. Consider Eq. (5a) with the leftmost carries c_0 and c_1 added:

$$\begin{array}{cccccccccc}
 & & 1. & b_1 & b_2 & \cdots & b_{p-3} & b_{p-2} & b_{p-1} & \times 2^{e_x} \\
\times & & 1. & 1 & & & & & & \times 2^{1} \\
\hline
 & & 1. & b_1 & b_2 & \cdots & b_{p-3} & b_{p-2} & b_{p-1} & \\
+ & 1. & b_1 & b_2 & \cdots & b_{p-3} & b_{p-2} & b_{p-1} & & \\
+ & c_0 & c_1 & \ldots & & & & & & \\
\hline
 & 1. & b'_1 & b'_2 & \cdots & b'_{p-3} & b'_{p-2} & b'_{p-1} & b'_p & \times 2^{e_x+1}
\end{array}$$

We have:

$$\begin{aligned}
1 + c_0 &= 1 \\
1 + b_1 + c_1 &= c_0 b'_1 \qquad \text{(henceforth, “}b_i b_j\text{” is to be interpreted as } b_i \times 2 + b_j\text{)}
\end{aligned}$$

Hence, $c_0 = 0$ and then $1 + b_1 + c_1 = b'_1$, which can hold only if $b_1 = c_1 = 0$ and $b'_1 = 1$. □

Lemma 2. *Given $x = 1.b_1 \ldots b_{p-1} \times 2^{e_x}$ a normal binary floating-point number. If $y = 3x$ has a $p+2$ bits significand, then $y = 1.0b'_2 \ldots b'_{p-1} b'_p b'_{p+1} \times 2^{e_x+2}$.*

Proof. Consider Eq. (5b) with the leftmost carries added:

$$\begin{array}{cccccccccc}
 & & 1. & b_1 & b_2 & \cdots & b_{p-3} & b_{p-2} & b_{p-1} & \times 2^{e_x} \\
\times & & 1. & 1 & & & & & & \times 2^{1} \\
\hline
 & & 1. & b_1 & b_2 & \cdots & b_{p-3} & b_{p-2} & b_{p-1} & \\
+ & 1. & b_1 & b_2 & \cdots & b_{p-3} & b_{p-2} & b_{p-1} & & \\
+ & c_0 & c_1 & & & & & & & \\
\hline
1. & b'_1 & b'_2 & \cdots & b'_{p-3} & b'_{p-2} & b'_{p-1} & b'_p & b'_{p+1} & \times 2^{e_x+2}
\end{array}$$

We have $1 + c_0 = 1b'_1$, from which we deduce $c_0 = 1$ and $b'_1 = 0$. □

When $y = 3x$ has a $p+2$ bits significand, we have to discard the last two bits to be able to represent the result in p bits format. If these two bits are “10”—that is, the real result is halfway between two representable numbers—, we need to know the antepenultimate bit of y's significand to decide the sign of the error. Consequently, we must consider the three leftmost bits of x. Appendix A presents all possible cases for the multiplication by 3.

For each possible value of x with respect to the five bits of its significand considered above (The implicit '1' bit before the radix point, the first bit to the right of the radix point, and the three rightmost bits), Tables 1 and 2 present all possible values for $\mathrm{fl}(3x)$. For homogeneity—required later on when we will investigate the expressions $3x/3$ and $3(x/3)$—, $\mathrm{fl}(3x)$ is also displayed with the same five bits, the exponents being left implicit. The third column “Error” gives the sign of the error:

$$\begin{cases}
\text{EQ}: & \mathrm{fl}(3x) = 3x \\
\text{GT}: & \mathrm{fl}(3x) > 3x \\
\text{LT}: & \mathrm{fl}(3x) < 3x
\end{cases}$$

3.2 The denormal case

Consider now the case in which $x = 0.b_1 \ldots b_{p-1} \times 2^{E_{\min}}$ is a denormal number. There are two cases depending on the value of b_1. For $b_1 = 0$, the following result holds:

Lemma 3. *Given a denormal floating-point number of the form $x = 0.0b_2 \ldots b_{p-1} \times 2^{E_{min}}$, we have:*

$$\mathrm{fl}(3x) = 3x \tag{6}$$

Table 1: Result and sign of the error for the multiplication by 3 of $x = 1.0b_2 \cdots b_{p-1} \times 2^{e_x}$

$\mathbf{x}$	fl($\mathbf{3x}$)	Error
$1.0b_2 \cdots b_{p-4}000$	$1.1b'_2 \cdots b'_{p-4}000$	EQ
	$1.1b'_2 \cdots b'_{p-4}100$	EQ
	$1.0b'_2 \cdots b'_{p-4}000$	EQ
	$1.0b'_2 \cdots b'_{p-4}010$	EQ
	$1.0b'_2 \cdots b'_{p-4}100$	EQ
	$1.0b'_2 \cdots b'_{p-4}110$	EQ
$1.0b_2 \cdots b_{p-4}001$	$1.1b'_2 \cdots b'_{p-4}010$	GT
	$1.1b'_2 \cdots b'_{p-4}110$	GT
	$1.0b'_2 \cdots b'_{p-4}001$	GT
	$1.0b'_2 \cdots b'_{p-4}011$	GT
	$1.0b'_2 \cdots b'_{p-4}101$	GT
	$1.0b'_2 \cdots b'_{p-4}111$	GT
$1.0b_2 \cdots b_{p-4}010$	$1.1b'_2 \cdots b'_{p-4}011$	EQ
	$1.1b'_2 \cdots b'_{p-4}111$	EQ
	$1.0b'_2 \cdots b'_{p-4}000$	GT
	$1.0b'_2 \cdots b'_{p-4}010$	GT
	$1.0b'_2 \cdots b'_{p-4}100$	GT
	$1.0b'_2 \cdots b'_{p-4}110$	GT
$1.0b_2 \cdots b_{p-4}011$	$1.1b'_2 \cdots b'_{p-4}000$	LT
	$1.1b'_2 \cdots b'_{p-4}100$	LT
	$1.0b'_2 \cdots b'_{p-4}000$	LT
	$1.0b'_2 \cdots b'_{p-4}010$	LT
	$1.0b'_2 \cdots b'_{p-4}100$	LT
	$1.0b'_2 \cdots b'_{p-4}110$	LT
$1.0b_2 \cdots b_{p-4}100$	$1.1b'_2 \cdots b'_{p-4}010$	EQ
	$1.1b'_2 \cdots b'_{p-4}110$	EQ
	$1.0b'_2 \cdots b'_{p-4}001$	EQ
	$1.0b'_2 \cdots b'_{p-4}011$	EQ
	$1.0b'_2 \cdots b'_{p-4}101$	EQ
	$1.0b'_2 \cdots b'_{p-4}111$	EQ
$1.0b_2 \cdots b_{p-4}101$	$1.1b'_2 \cdots b'_{p-4}000$	GT
	$1.1b'_2 \cdots b'_{p-4}100$	GT
	$1.0b'_2 \cdots b'_{p-4}000$	GT
	$1.0b'_2 \cdots b'_{p-4}010$	GT
	$1.0b'_2 \cdots b'_{p-4}100$	GT
	$1.0b'_2 \cdots b'_{p-4}110$	GT
$1.0b_2 \cdots b_{p-4}110$	$1.1b'_2 \cdots b'_{p-4}001$	EQ
	$1.1b'_2 \cdots b'_{p-4}101$	EQ
	$1.0b'_2 \cdots b'_{p-4}000$	LT
	$1.0b'_2 \cdots b'_{p-4}010$	LT
	$1.0b'_2 \cdots b'_{p-4}100$	LT
	$1.0b'_2 \cdots b'_{p-4}110$	LT
$1.0b_2 \cdots b_{p-4}111$	$1.1b'_2 \cdots b'_{p-4}010$	LT
	$1.1b'_2 \cdots b'_{p-4}110$	LT
	$1.0b'_2 \cdots b'_{p-4}001$	LT
	$1.0b'_2 \cdots b'_{p-4}011$	LT
	$1.0b'_2 \cdots b'_{p-4}101$	LT
	$1.0b'_2 \cdots b'_{p-4}111$	LT

Table 2: Result and sign of the error for the multiplication by 3 of $x = 1.1b_2 \cdots b_{p-1} \times 2^{e_x}$

$\mathbf{x}$	fl($\mathbf{3x}$)	Error
$1.1b_2 \cdots b_{p-4}000$	$1.0b'_2 \cdots b'_{p-4}000$	EQ
	$1.0b'_2 \cdots b'_{p-4}010$	EQ
	$1.0b'_2 \cdots b'_{p-4}100$	EQ
	$1.0b'_2 \cdots b'_{p-4}110$	EQ
$1.1b_2 \cdots b_{p-4}001$	$1.0b'_2 \cdots b'_{p-4}001$	GT
	$1.0b'_2 \cdots b'_{p-4}011$	GT
	$1.0b'_2 \cdots b'_{p-4}101$	GT
	$1.0b'_2 \cdots b'_{p-4}111$	GT
$1.1b_2 \cdots b_{p-4}010$	$1.0b'_2 \cdots b'_{p-4}000$	GT
	$1.0b'_2 \cdots b'_{p-4}010$	GT
	$1.0b'_2 \cdots b'_{p-4}100$	GT
	$1.0b'_2 \cdots b'_{p-4}110$	GT
$1.1b_2 \cdots b_{p-4}011$	$1.0b'_2 \cdots b'_{p-4}000$	LT
	$1.0b'_2 \cdots b'_{p-4}010$	LT
	$1.0b'_2 \cdots b'_{p-4}100$	LT
	$1.0b'_2 \cdots b'_{p-4}110$	LT
$1.1b_2 \cdots b_{p-4}100$	$1.0b'_2 \cdots b'_{p-4}001$	EQ
	$1.0b'_2 \cdots b'_{p-4}011$	EQ
	$1.0b'_2 \cdots b'_{p-4}101$	EQ
	$1.0b'_2 \cdots b'_{p-4}111$	EQ
$1.1b_2 \cdots b_{p-4}101$	$1.0b'_2 \cdots b'_{p-4}000$	GT
	$1.0b'_2 \cdots b'_{p-4}010$	GT
	$1.0b'_2 \cdots b'_{p-4}100$	GT
	$1.0b'_2 \cdots b'_{p-4}110$	GT
$1.1b_2 \cdots b_{p-4}110$	$1.0b'_2 \cdots b'_{p-4}000$	LT
	$1.0b'_2 \cdots b'_{p-4}010$	LT
	$1.0b'_2 \cdots b'_{p-4}100$	LT
	$1.0b'_2 \cdots b'_{p-4}110$	LT
$1.1b_2 \cdots b_{p-4}111$	$1.0b'_2 \cdots b'_{p-4}001$	LT
	$1.0b'_2 \cdots b'_{p-4}011$	LT
	$1.0b'_2 \cdots b'_{p-4}101$	LT
	$1.0b'_2 \cdots b'_{p-4}111$	LT

Proof. Trivial. Consider:

$$\begin{array}{cccccccccc}
 & & 0. & 0 & b_2 & & \cdots & & b_{p-1} & \times 2^{E_{\min}} \\
\times & & 1. & 1 & & & & & & \times 2^{1} \\
\hline
 & & 0. & 0 & b_2 & b_3 & \cdots & b_{p-1} & & \\
+ & 0. & 0 & b_2 & b_3 & \cdots & b_{p-1} & & & \\
\hline
 & 0 & b'_1. & b'_2 & b'_3 & \cdots & b'_{p-1} & b'_p & & \times 2^{E_{\min}}
\end{array} \tag{7}$$

□

On the other hand, if $b_1 \neq 0$, we have:

$$\begin{array}{cccccccccc}
 & & 0. & 1 & b_2 & \cdots & b_{p-1} & & & \times 2^{E_{\min}} \\
\times & & 1. & 1 & & & & & & \times 2^{1} \\
\hline
 & & 0. & 1 & b_2 & & \cdots & & b_{p-1} & \\
+ & 0. & 1 & b_2 & & \cdots & & b_{p-1} & & \\
 & c_0 & c_1 & & & & & & & \\
\hline
 & b'_0. & b'_1 & b'_2 & b'_3 & \cdots & b'_{p-1} & b'_p & & \times 2^{E_{\min}+1}
\end{array}$$

where c_0 and c_1 are, as before, the leftmost carries.

It is easy to see that c_0 and c_1 must be equal. Furthermore:

$$\begin{cases} \text{if } c_0 = c_1 = 0, & \text{then } b'_0 = 0 \text{ and } b'_1 = 1 \\ \text{if } c_0 = c_1 = 1, & \text{then } b'_0 = 1 \text{ and } b'_1 = 0 \end{cases}$$

Multiplying x by 3 can then lead to a p bits significand (when $b'_0 = 0$) or to a $p+1$ bits significand (when $b'_0 = 1$). Appendix B displays the different results for all possible values of the rightmost bits[2].

From Appendix B and Eq. (7), we can deduce the value of $\mathrm{fl}(3x)$ obtained from rounding $3x$ to nearest-even (see Table 3).

4 Dividing x by three

Similarly to the multiplication, we will first consider the case in which x is a normal number and no underflow occurs; we will then investigate what happens when x is a denormal number or when the result underflows.

For $x = (-1)^{s_x} m_x \times 2^{e_x}$ and $y = (-1)^{s_y} m_y \times 2^{e_y}$, the quotient x/y is given by [9]:

$$x/y = (-1)^{s_x \oplus s_y} m_x/m_y \times 2^{e_x - e_y} \tag{8}$$

To determine the sign of the error of $\mathrm{fl}(x/y)$ relative to x/y, we may only consider the significands m_x and m_y. Consequently, we will only display the exponents when rendered necessary by the context. We also restrict ourselves to positive values for x, the negative case being easily deduced.

4.1 The normal case

As for the multiplication, we need only consider the first bit after the radix point and the last three bits of the significand to determine the result of the division by 3. The following three lemmas tell us how many bits to compute in order to obtain a p bits significand correctly rounded to nearest-even.

Lemma 4. *If $x = 1.1b_2 \dots b_{p-1} \times 2^{e_x}$ is a normal floating-point number, then we need to compute $p+1$ bits to get a correctly rounded p bits significand for $y = \mathrm{fl}(x/3)$; in addition, we have:*

$$\mathrm{fl}(y) = 1.0b'_2 \dots b'_{p-1} \times 2^{e_x - 1} \tag{9}$$

provided no underflow occurs.

[2]Note that, even though only the last two bits need to be considered, we take into account the last three as in the previous section.

Table 3: Result and sign of the error for the multiplication by 3 of $x = 0.b_1 b_2 \cdots b_{p-1} \times 2^{E_{\min}}$

$\mathbf{x}$	fl($\mathbf{3x}$)	Error
$0.0b_2 \cdots b_{p-1}$	$b'_0.b'_1 \cdots b'_{p-1}$	EQ
$0.1b_2 \cdots b_{p-4}000$	$1.1b'_2 \cdots b'_{p-4}000$	EQ
	$1.0b'_2 \cdots b'_{p-3}00$	EQ
$0.1b_2 \cdots b_{p-4}001$	$1.1b'_2 \cdots b'_{p-4}011$	EQ
	$1.0b'_2 \cdots b'_{p-3}10$	GT
$0.1b_2 \cdots b_{p-4}010$	$1.1b'_2 \cdots b'_{p-4}110$	EQ
	$1.0b'_2 \cdots b'_{p-3}110$	EQ
$0.1b_2 \cdots b_{p-4}011$	$1.1b'_2 \cdots b'_{p-4}001$	EQ
	$1.0b'_2 \cdots b'_{p-3}00$	LT
$0.1b_2 \cdots b_{p-4}100$	$1.1b'_2 \cdots b'_{p-4}100$	EQ
	$1.0b'_2 \cdots b'_{p-3}10$	EQ
$0.1b_2 \cdots b_{p-4}101$	$1.1b'_2 \cdots b'_{p-4}111$	EQ
	$1.0b'_2 \cdots b'_{p-3}00$	GT
$0.1b_2 \cdots b_{p-4}110$	$1.1b'_2 \cdots b'_{p-4}010$	EQ
	$1.0b'_2 \cdots b'_{p-3}01$	EQ
$0.1b_2 \cdots b_{p-4}111$	$1.1b'_2 \cdots b'_{p-4}101$	EQ
	$1.0b'_2 \cdots b'_{p-3}10$	LT

Proof. Consider the process of dividing x by 3 at its beginning:

$$\begin{array}{rccccc} & \mathit{0} & 1 & 0 & \cdots & \\ \cline{2-6} 11) & 1 & 1 & b_2 & \ldots & b_{p-1} \\ & 1 & 1 & & & \\ \cline{2-3} & & 0 & b_2 & & \\ & & & \cdots & & \end{array}$$

Due to the leading 0 of the quotient, we need to compute one bit more to get p bits in the rounded significand. Equation (9) is obvious from the division process shown above. □

Lemma 5. *If $x = 1.0b_2 \ldots b_{p-1} \times 2^{e_x}$ is a normal floating-point number, then we need to compute $p + 2$ bits to get a p bit significand for $y = x/3$.*

Proof. Consider the process of dividing x by 3 at its beginning:

$$\begin{array}{rccccccc} & \mathit{0} & \mathit{0} & 1 & b'_1 & \cdots & & \\ \cline{2-7} 11) & 1 & 0 & b_2 & b_3 & \ldots & b_{p-1} \\ & 1 & 0 & b_2 & & & \\ \cline{2-4} & r_0 & & r_1 & b_3 & & \\ & & & \cdots & & & \end{array}$$

Due to the two leading 0 of the quotient, we need to compute two bits more to get p bits in the significand of the result. □

Lemma 6. *Given $x = 1.b_1 \ldots b_{p-1} \times 2^{e_x}$ a normal floating-point number and $y = x/3$. There are three possible forms for the significand of y:*

$$\begin{cases} 1.b'_1 \ldots b'_{p-2}\overline{0} \\ 1.b'_1 \ldots b'_{p-2}\overline{01} \\ 1.b'_1 \ldots b'_{p-2}\overline{10} \end{cases}$$

where $\overline{s}$ denotes an infinite repetition of the binary string s.

Proof. In the process of dividing x by 3, there are only three possible remainders at each step: 0_2, 1_2, or 10_2. Consequently, when $x = 1.1b_2 \dots b_{p-1} \times 2^{e_x}$, we only have the following three situations (the case $x = 1.0b_2 \dots b_{p-1} \times 2^{e_x}$ is analogous):

$$
\begin{array}{rcccccccc}
 & \mathbf{0} & 1 & 0 & b'_2 & \cdots & b'_{p-2} & \overline{0} & \dots \\
11) & 1 & 1 & b_2 & & \dots & b_{p-1} & & \\
 & & & & \dots & & & \downarrow & \\
 & & & & & & \boxed{0} & 0 &
\end{array}
\qquad
\begin{array}{rcccccccc}
 & \mathbf{0} & 1 & 0 & b'_2 & \cdots & b'_{p-2} & \overline{0} & 1 & \dots \\
11) & 1 & 1 & b_2 & & \dots & b_{p-1} & & & \\
 & & & & & \dots & & \downarrow & \downarrow & \\
 & & & & & & \boxed{1} & 0 & 0 & \\
 & & & & & & & & 1 &
\end{array}
$$

$$
\begin{array}{rcccccccc}
 & \mathbf{0} & 1 & 0 & b'_2 & \cdots & b'_{p-2} & 1 & 0 & \dots \\
11) & 1 & 1 & b_2 & & \dots & b_{p-1} & & & \\
 & & & & \dots & & & \downarrow & & \\
 & & & & & & \boxed{10} & 0 & \downarrow & \\
 & & & & & & & 1 & 0 &
\end{array}
\tag{10}
$$

□

For each triplet of rightmost bits of the significand of x, there are three cases to consider depending on the remainder after having taken into account the bit b_{p-4}. All the cases are presented in Section C.

Tables 4 and 5 summarize the results for the whole section.

4.2 The denormal case

There are two possibilities for an underflow to occur when dividing x by 3: either x is already a denormal number, or x is normal but the rounded result of the division is a denormal number.

If x is a denormal number, we have:

$$
\begin{array}{rccccccc}
 & 0. & 0 & b'_2 & & \cdots & & b'_{p-1} \times 2^{E_{\min}} \\
1.1 \times 2^1) & 0. & b_1 & & \dots & b_{p-4} \; b_{p-3} \; b_{p-2} & & b_{p-1} \times 2^{E_{\min}} \\
 & & \dots & & & & & \\
 & & & \dots & & & & \\
 & & & & \dots & & & \\
 & & & & & \dots & &
\end{array}
\tag{11}
$$

If x is a normal number of the form $x = 1.b_1 \dots b_{p-1} \times 2^{E_{\min}}$, we have, either:

$$
\begin{array}{rccccccc}
 & 0. & 0 & b'_2 & & \cdots & & b'_{p-1} \times 2^{E_{\min}} \\
1.1 \times 2^1) & 1. & 0 & & \dots & b_{p-4} \; b_{p-3} \; b_{p-2} & & b_{p-1} \times 2^{E_{\min}} \\
 & & \dots & & & & & \\
 & & & \dots & & & & \\
 & & & & \dots & & & \\
 & & & & & \dots & &
\end{array}
\tag{12}
$$

or:

$$
\begin{array}{rccccccc}
 & 0. & 1 & b'_2 & & \cdots & & b'_{p-1} \times 2^{E_{\min}} \\
1.1 \times 2^1) & 1. & 1 & & \dots & b_{p-4} \; b_{p-3} \; b_{p-2} & & b_{p-1} \times 2^{E_{\min}} \\
 & & \dots & & & & & \\
 & & & \dots & & & & \\
 & & & & \dots & & & \\
 & & & & & \dots & &
\end{array}
\tag{13}
$$

In all three cases, the result of the division cannot be scaled to the right since the exponent is the smallest possible. As a consequence, we need only compute p bits to obtain $\mathrm{fl}(x/3) = 0.b'_1 \dots b'_{p-3}b'_{p-2}b'_{p-1} \times 2^{E_{\min}}$.

Table 4: Result and sign of the error for the division by 3 of $x = 1.0b_2 \cdots b_{p-1} \times 2^{e_x}$ with no underflow

$\mathbf{x}$	fl($\mathbf{x}/\mathbf{3}$)	Error
$1.0b_2 \cdots b_{p-4}000$	$1.0b'_2 \cdots b'_{p-4}000$	EQ
	$1.1b'_2 \cdots b'_{p-4}000$	EQ
	$1.0b'_2 \cdots b'_{p-4}011$	GT
	$1.1b'_2 \cdots b'_{p-4}011$	GT
	$1.0b'_2 \cdots b'_{p-4}101$	LT
	$1.1b'_2 \cdots b'_{p-4}101$	LT
$1.0b_2 \cdots b_{p-4}001$	$1.0b'_2 \cdots b'_{p-4}001$	LT
	$1.1b'_2 \cdots b'_{p-4}001$	LT
	$1.0b'_2 \cdots b'_{p-4}100$	EQ
	$1.1b'_2 \cdots b'_{p-4}100$	EQ
	$1.0b'_2 \cdots b'_{p-4}111$	GT
	$1.1b'_2 \cdots b'_{p-4}111$	GT
$1.0b_2 \cdots b_{p-4}010$	$1.0b'_2 \cdots b'_{p-4}011$	GT
	$1.1b'_2 \cdots b'_{p-4}011$	GT
	$1.0b'_2 \cdots b'_{p-4}000$	EQ
	$1.1b'_2 \cdots b'_{p-4}000$	EQ
	$1.0b'_2 \cdots b'_{p-4}101$	LT
	$1.1b'_2 \cdots b'_{p-4}101$	LT
$1.0b_2 \cdots b_{p-4}011$	$1.0b'_2 \cdots b'_{p-4}100$	EQ
	$1.1b'_2 \cdots b'_{p-4}100$	EQ
	$1.0b'_2 \cdots b'_{p-4}111$	GT
	$1.1b'_2 \cdots b'_{p-4}111$	GT
	$1.0b'_2 \cdots b'_{p-4}001$	LT
	$1.1b'_2 \cdots b'_{p-4}001$	LT
$1.0b_2 \cdots b_{p-4}100$	$1.0b'_2 \cdots b'_{p-4}101$	LT
	$1.1b'_2 \cdots b'_{p-4}101$	LT
	$1.0b'_2 \cdots b'_{p-4}000$	EQ
	$1.1b'_2 \cdots b'_{p-4}000$	EQ
	$1.0b'_2 \cdots b'_{p-4}011$	GT
	$1.1b'_2 \cdots b'_{p-4}011$	GT
$1.0b_2 \cdots b_{p-4}101$	$1.0b'_2 \cdots b'_{p-4}111$	GT
	$1.1b'_2 \cdots b'_{p-4}111$	GT
	$1.0b'_2 \cdots b'_{p-4}001$	LT
	$1.1b'_2 \cdots b'_{p-4}001$	LT
	$1.0b'_2 \cdots b'_{p-4}100$	EQ
	$1.1b'_2 \cdots b'_{p-4}100$	EQ
$1.0b_2 \cdots b_{p-4}110$	$1.0b'_2 \cdots b'_{p-4}000$	EQ
	$1.1b'_2 \cdots b'_{p-4}000$	EQ
	$1.0b'_2 \cdots b'_{p-4}011$	GT
	$1.1b'_2 \cdots b'_{p-4}011$	GT
	$1.0b'_2 \cdots b'_{p-4}101$	LT
	$1.1b'_2 \cdots b'_{p-4}101$	LT
$1.0b_2 \cdots b_{p-4}111$	$1.0b'_2 \cdots b'_{p-4}001$	LT
	$1.1b'_2 \cdots b'_{p-4}001$	LT
	$1.0b'_2 \cdots b'_{p-4}100$	EQ
	$1.1b'_2 \cdots b'_{p-4}100$	EQ
	$1.0b'_2 \cdots b'_{p-4}111$	GT
	$1.1b'_2 \cdots b'_{p-4}111$	GT

Table 5: Result and sign of the error for the division by 3 of $x = 1.1b_2 \cdots b_{p-1} \times 2^{e_x}$ with no underflow.

$\mathbf{x}$	fl($\mathbf{x}/\mathbf{3}$)	Error
$1.1b_2 \cdots b_{p-4}000$	$1.0b'_2 \cdots b'_{p-4}000$	EQ
	$1.0b'_2 \cdots b'_{p-4}101$	LT
	$1.0b'_2 \cdots b'_{p-4}011$	GT
$1.1b_2 \cdots b_{p-4}001$	$1.0b'_2 \cdots b'_{p-4}001$	GT
	$1.0b'_2 \cdots b'_{p-4}110$	EQ
	$1.0b'_2 \cdots b'_{p-4}011$	LT
$1.1b_2 \cdots b_{p-4}010$	$1.0b'_2 \cdots b'_{p-4}001$	LT
	$1.0b'_2 \cdots b'_{p-4}100$	EQ
	$1.0b'_2 \cdots b'_{p-4}111$	GT
$1.1b_2 \cdots b_{p-4}011$	$1.0b'_2 \cdots b'_{p-4}010$	EQ
	$1.0b'_2 \cdots b'_{p-4}111$	LT
	$1.0b'_2 \cdots b'_{p-4}101$	GT
$1.1b_2 \cdots b_{p-4}100$	$1.0b'_2 \cdots b'_{p-4}011$	GT
	$1.0b'_2 \cdots b'_{p-4}000$	EQ
	$1.0b'_2 \cdots b'_{p-4}101$	LT
$1.1b_2 \cdots b_{p-4}101$	$1.0b'_2 \cdots b'_{p-4}011$	LT
	$1.0b'_2 \cdots b'_{p-4}001$	GT
	$1.0b'_2 \cdots b'_{p-4}110$	EQ
$1.1b_2 \cdots b_{p-4}110$	$1.0b'_2 \cdots b'_{p-4}100$	EQ
	$1.0b'_2 \cdots b'_{p-4}001$	LT
	$1.0b'_2 \cdots b'_{p-4}111$	GT
$1.1b_2 \cdots b_{p-4}111$	$1.0b'_2 \cdots b'_{p-4}101$	GT
	$1.0b'_2 \cdots b'_{p-4}010$	EQ
	$1.0b'_2 \cdots b'_{p-4}111$	LT

One can easily check that, for $x = 0.b_1 \cdots b_{p-1} \times 2^{E_{\min}}$ and $x = 1.0b_2 \cdots b_{p-1} \times 2^{E_{\min}}$, the bits b'_{p-3}, b'_{p-2}, b'_{p-1} must be equal to, respectively, b'_{p-5}, b'_{p-4}, b'_{p-3} as obtained when dividing a normal number of the form $1.0b_2 \ldots b_{p-1} \times 2^e$—with $e > E_{\min}$—by 3. When $x = 1.1b_2 \cdots b_{p-1} \times 2^{E_{\min}}$, the bits b'_{p-3}, b'_{p-2}, b'_{p-1} must be equal to, respectively, b'_{p-4}, b'_{p-3}, and b'_{p-2} when dividing by 3 a normal number of the form $x = 1.1b_2 \cdots b_{p-1} \times 2^e$, with $e > E_{\min}$.

For an underflow to occur when dividing a normal number x by 3, x must be smaller or equal to $(3\lambda)^-$—with $(3\lambda)^- = 1.01 \cdots 111 \times 2^{E_{\min}+1}$—, where $\lambda = 1.0 \times 2^{E_{\min}}$ is the smallest positive normal floating-point number. The case in which the exponent of x is $E_{\min}$ has already been dealt with above. When x is of the form $1.0b_2 \ldots b_{p-1} \times 2^{E_{\min}+1}$, it is easy to see in Eq. (14):

$$
\begin{array}{rcccccccccc}
 & 0 & 0. & 1 & & \cdots & & b'_{p-3} & b'_{p-2} & b'_{p-1} & \times 2^{E_{\min}} \\
1.1 \times 2^1) & 1. & 0 & b_2 & \ldots & b_{p-4} & b_{p-3} & b_{p-2} & b_{p-1} & & \times 2^{E_{\min+1}} \\
 & & \ldots & & & & & & & & \\
 & & & \ldots & & & & & & & \\
 & & & & & \ldots & & & & & \\
 & & & & & & & \ldots & & &
\end{array}
\tag{14}
$$

that we need to compute $p+1$ bits in order to get a p bits significand (barring the rounding process), and that the last three bits of the result are obtained following the same rules as for the case in which x is of the form $1.1b_2 \ldots b_p \times 2^e$ with $e > E_{\min} + 1$.

Table 6: Result and sign of the error for the division by 3 of $x = 1.b_1 \cdots b_{p-1} \times 2^{E_{\min}}$ or $x = 0.b_1 \ldots b_{p-1} \times 2^{E_{\min}}$ with underflow. As shown in Eqs. (11), (12) and (13), $b'_1 = 1$ if and only if $x = 1.1b_2 \cdots b_{p-1} \times 2^{E_{\min}}$

x	fl(**x**/**3**)	Error
$b_0.b_1b_2 \cdots b_{p-4}000$	$0.b'_1 \cdots b'_{p-4}000$	EQ
	$0.b'_1 \cdots b'_{p-4}011$	GT
	$0.b'_1 \cdots b'_{p-4}101$	LT
$b_0.b_1b_2 \cdots b_{p-4}001$	$0.b'_1 \cdots b'_{p-4}000$	LT
	$0.b'_1 \cdots b'_{p-4}011$	EQ
	$0.b'_1 \cdots b'_{p-4}110$	GT
$b_0.b_1b_2 \cdots b_{p-4}010$	$0.b'_1 \cdots b'_{p-4}001$	GT
	$0.b'_1 \cdots b'_{p-4}110$	EQ
	$0.b'_1 \cdots b'_{p-4}011$	LT
$b_0.b_1b_2 \cdots b_{p-4}011$	$0.b'_1 \cdots b'_{p-4}001$	EQ
	$0.b'_1 \cdots b'_{p-4}100$	GT
	$0.b'_1 \cdots b'_{p-4}110$	LT
$b_0.b_1b_2 \cdots b_{p-4}100$	$0.b'_1 \cdots b'_{p-4}001$	LT
	$0.b'_1 \cdots b'_{p-4}100$	EQ
	$0.b'_1 \cdots b'_{p-4}111$	GT
$b_0.b_1b_2 \cdots b_{p-4}101$	$0.b'_1 \cdots b'_{p-4}010$	GT
	$0.b'_1 \cdots b'_{p-4}100$	LT
	$0.b'_1 \cdots b'_{p-4}111$	EQ
$b_0.b_1b_2 \cdots b_{p-4}110$	$0.b'_1 \cdots b'_{p-4}010$	EQ
	$0.b'_1 \cdots b'_{p-4}101$	GT
	$0.b'_1 \cdots b'_{p-4}111$	LT
$b_0.b_1b_2 \cdots b_{p-4}111$	$0.b'_1 \cdots b'_{p-4}010$	LT
	$0.b'_1 \cdots b'_{p-4}101$	EQ
	$0.b'_1 \cdots b'_{p-4}000$	GT

Tables 6 and 7 summarize all the previous results for the denormal case.

Table 7: Result and sign of the error for the division by 3 of $x = 1.0b_2 \cdots b_{p-1} \times 2^{E_{\min}+1}$ with underflow

$\mathbf{x}$	fl($\mathbf{x}/\mathbf{3}$)	Error
$1.0b_2 \cdots b_{p-4}000$	$0.1 \cdots b'_{p-4}000$	EQ
	$0.1 \cdots b'_{p-4}101$	LT
	$0.1 \cdots b'_{p-4}011$	GT
$1.0b_2 \cdots b_{p-4}001$	$0.1 \cdots b'_{p-4}001$	GT
	$0.1 \cdots b'_{p-4}110$	EQ
	$0.1 \cdots b'_{p-4}011$	LT
$1.0b_2 \cdots b_{p-4}010$	$0.1 \cdots b'_{p-4}001$	LT
	$0.1 \cdots b'_{p-4}100$	EQ
	$0.1 \cdots b'_{p-4}111$	GT
$1.0b_2 \cdots b_{p-4}011$	$0.1 \cdots b'_{p-4}010$	EQ
	$0.1 \cdots b'_{p-4}111$	LT
	$0.1 \cdots b'_{p-4}101$	GT
$1.0b_2 \cdots b_{p-4}100$	$0.1 \cdots b'_{p-4}011$	GT
	$0.1 \cdots b'_{p-4}000$	EQ
	$0.1 \cdots b'_{p-4}101$	LT
$1.0b_2 \cdots b_{p-4}101$	$0.1 \cdots b'_{p-4}011$	LT
	$0.1 \cdots b'_{p-4}001$	GT
	$0.1 \cdots b'_{p-4}110$	EQ
$1.0b_2 \cdots b_{p-4}110$	$0.1 \cdots b'_{p-4}100$	EQ
	$0.1 \cdots b'_{p-4}001$	LT
	$0.1 \cdots b'_{p-4}111$	GT
$1.0b_2 \cdots b_{p-4}111$	$0.1 \cdots b'_{p-4}101$	GT
	$0.1 \cdots b'_{p-4}010$	EQ
	$0.1 \cdots b'_{p-4}111$	LT

5 Composing the multiplication and division

As seen in Section 2, the correctness of the floating-point multiplication and division are precisely defined by IEEE 754. From Eq. (2) and Eq. (3), we may compute the worst case error for the composition of a multiplication and a division. In theory, there are eight cases depending on the order of the operations and on the outcome of each operation (underflow or not). Tables 8 and 9 summarize the error analysis when computing either $\mathrm{fl}\langle(3x)/3\rangle$ or $\mathrm{fl}\langle 3(x/3)\rangle$. The crossed out cells in the tables correspond to cases that cannot occur: Consider, for example, Table 8; if there is an underflow when multiplying x by 3, there will surely be an underflow when dividing $3x$ by 3.

Table 8: Worst case error analysis for $\mathrm{fl}\langle\frac{3x}{3}\rangle$, with: $|\eta_1| \leqslant \mu/2$, $|\eta_2| \leqslant \mu/2$, $|\delta_1| \leqslant u/(1+u)$, $|\delta_2| \leqslant u - 2u^2$.

$y = 3x$ \ $z = y/3$	Underflow	No underflow
Underflow	$\frac{3x+\eta_1}{3} + \eta_2$	~~$\frac{3x+\eta_1}{3}(1+\delta_2)$~~
No underflow	$\frac{3x(1+\delta_1)}{3} + \eta_2$	$\frac{3x}{3}(1+\delta_1)(1+\delta_2)$

Table 9: Worst case error analysis for $\mathrm{fl}\langle 3\frac{x}{3}\rangle$, with: $|\eta_1| \leqslant \mu/2$, $|\eta_2| \leqslant \mu/2$, $|\delta_1| \leqslant u - 2u^2$, $|\delta_2| \leqslant u/(1+u)$.

$y = x/3$ \ $z = 3y$	Underflow	No underflow
Underflow	$3\left(\frac{x}{3}+\eta_1\right) + \eta_2$	$3\left(\frac{x}{3}+\eta_1\right)(1+\delta_2)$
No underflow	~~$3\left(\frac{x}{3}(1+\delta_1)\right) + \eta_2$~~	$3\frac{x}{3}(1+\delta_1)(1+\delta_2)$

5.1 Multiplication and division with no underflow

Considering first the case in which no underflow occurs, we get, irrespective of the order of the operations considered:

$$\mathrm{fl}\langle\frac{3x}{3}\rangle = \mathrm{fl}\langle 3\frac{x}{3}\rangle = x(1+\delta_a)(1+\delta_b)$$

with $|\delta_a| \leqslant u/(1+u)$ and $|\delta_b| \leqslant u - 2u^2$. Equivalently, we have:

$$\mathrm{fl}\langle\frac{3x}{3}\rangle = \mathrm{fl}\langle 3\frac{x}{3}\rangle = x(1+\delta) \tag{15}$$

with $\delta = \delta_a + \delta_b + \delta_a\delta_b$, which means that $|\delta| \leqslant 2(u-2u^3)/(1+u)$, that is $|\delta| < 2u$. The relative distance between two consecutive floating-point numbers is contained in the range $[u, 2u]$; Consequently, the rounded value of $3x/3$ or $3(x/3)$ may not round to x exactly, but we have the guarantee that it must be at worst one of the immediate floating-point neighbors of x (See Fig. 1).

5.2 Multiplication and division with underflow

As can be seen in Table 8 and Table 9, the worst case errors are different for $\mathrm{fl}\langle(3x)/3\rangle$ and $\mathrm{fl}\langle 3(x/3)\rangle$ whenever an underflow occurs. We must, therefore, consider separately the two computations.

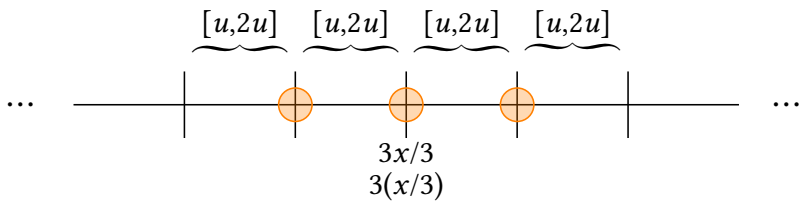


Figure 1: Orange disks representing the possible values for $\mathrm{fl}\langle 3x/3\rangle$ or $\mathrm{fl}\langle 3(x/3)\rangle$ with rounding to nearest-even in the absence of underflow.

5.2.1 The case of $\mathrm{fl}\langle(3x)/3\rangle$

Underflow on multiplication and division. When both the multiplication and division underflow, we get:

$$\mathrm{fl}\langle\frac{3x}{3}\rangle = \frac{3x+\eta_1}{3} + \eta_2 = x + \left(\frac{\eta_1}{3} + \eta_2\right), \text{ with } |\eta_1| \leqslant \frac{\mu}{2}, |\eta_2| \leqslant \frac{\mu}{2}$$

Hence,

$$\mathrm{fl}\langle\frac{3x}{3}\rangle = x + \eta, \text{ with } |\eta| \leqslant \frac{2\mu}{3} \tag{16}$$

For the multiplication to underflow, x must be a subnormal number. As shown in Figure 2, the absolute distance between two consecutive subnormals is μ. Consequently, we deduce from Eq. (16) that

$$\mathrm{fl}\langle\frac{3x}{3}\rangle = \frac{3x}{3} = x$$

if both the multiplication and the division underflow.

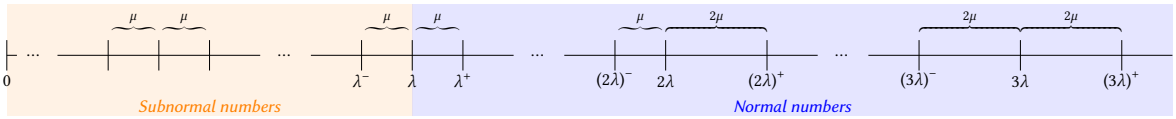


Figure 2: Absolute distance between small positive floating-point numbers

Underflow on division but not on multiplication. If the multiplication does not underflow but the division does, we get:

$$\mathrm{fl}\langle\frac{3x}{3}\rangle = \frac{3x(1+\delta_1)}{3} + \eta_2, \text{ with } |\delta_1| \leqslant \frac{u}{1+u}, |\eta_2| \leqslant \frac{\mu}{2}.$$

It is easy to check that for $x = \lambda$, we have $\mathrm{fl}\langle(3x)/3\rangle = \lambda$. By monotonicity of rounding, we get that for all y greater than x, we have $\mathrm{fl}\langle(3y)/3\rangle \geqslant \lambda$. Hence, x must be a subnormal number for the division to underflow when computing $(3x)/3$. Therefore, the absolute distance between $3x$ and $\mathrm{fl}(3x)$ must be smaller or equal to μ (See Figure 2), which leads to:

$$\begin{aligned}\mathrm{fl}\langle\frac{3x}{3}\rangle &= \frac{3x+\eta_3}{3} + \eta_2, \text{ with } |\eta_3| \leqslant \mu, |\eta_2| \leqslant \frac{\mu}{2} \\ &= x + \frac{5}{6}\mu\end{aligned}$$

As in the previous case, we deduce:

$$\mathrm{fl}\langle\frac{3x}{3}\rangle = \frac{3x}{3} = x$$

if the multiplication does not underflow but the division does.

5.2.2 The case of $\mathrm{fl}\langle 3(x/3)\rangle$

Underflow on division and multiplication. If both the division and the multiplication underflow, we get:

$$\mathrm{fl}\langle 3\frac{x}{3}\rangle = 3\left(\frac{x}{3} + \eta_1\right) + \eta_2, \text{ with } |\eta_1| \leqslant \mu/2, |\eta_2| \leqslant \mu/2 \tag{17}$$

Using monotonicity of rounding, if $\mathrm{fl}\langle 3(x/3)\rangle$ underflows, we must have $x \leqslant \lambda$. Therefore, $\mathrm{fl}(x/3)$ is of the form $0.0b_2' \cdots b_{p-1}' \times 2^{E_{\min}}$. According to Table 3, the multiplication is then without error, that is: $\eta_2 = 0$. Equation (17) simplifies to:

$$\mathrm{fl}\langle 3\frac{x}{3}\rangle = x + 3\eta_1, \text{ with } |\eta_1| \leqslant \mu/2$$

Hence,

$$|\mathrm{fl}\langle 3\frac{x}{3}\rangle - x| \leqslant \frac{3\mu}{2}$$

Since the absolute distance between subnormal numbers is μ, we have:

$$\mathrm{fl}\langle 3\frac{x}{3}\rangle \in \{x^-, x, x^+\}$$

Underflow on division but not on multiplication. If the division underflows but the multiplication does not, we get:

$$\mathrm{fl}\langle 3\frac{x}{3}\rangle = 3\left(\frac{x}{3} + \eta_1\right)(1 + \delta_2), \text{ with } |\eta_1| \leqslant \mu/2, |\delta_2| \leqslant u/(1+u) \tag{18}$$

It is trivial to check that we must have $x \in [\lambda^+, (3\lambda)^-]$, using monotonicity of rounding. For convenience, we will investigate the actual worst case error on three disjoint domains for x:

- If $x \in [\lambda^+, (\frac{3}{2}\lambda)^-]$ (i.e., $x \in [1.00\cdots001 \times 2^{E_{\min}}, 1.011\cdots111 \times 2^{E_{\min}}]$).

  According to Table 6, we have:

  $$x \in [\lambda^+, (\frac{3}{2}\lambda)^-] \implies \mathrm{fl}\left(\frac{x}{3}\right) = 0.0b_2' \cdots b_{p-1}' \times 2^{E_{\min}}$$

  According to Table 3, we therefore have:

  $$\mathrm{fl}\left(3\mathrm{fl}\left(\frac{x}{3}\right)\right) = 3\mathrm{fl}\left(\frac{x}{3}\right)$$

  which means:

  $$\mathrm{fl}\langle 3\frac{x}{3}\rangle = 3(\frac{x}{3} + \eta_1), |\eta_1| \leqslant \mu/2$$

  Therefore:

  $$|\mathrm{fl}\langle 3\frac{x}{3}\rangle - x| \leqslant \frac{3\mu}{2}$$

  Since, the absolute distance between floating-point numbers is μ in the domain $[\lambda^+, (\frac{3}{2}\lambda)^-]$, we have:

  $$\forall x \in [\lambda^+, (3\lambda/2)^-] : \mathrm{fl}\langle 3\frac{x}{3}\rangle \in \{x^-, x, x^+\} \tag{19}$$

- If $x \in [3\lambda/2, (2\lambda)^-]$ (i.e. $x \in [1.10\cdots000 \times 2^{E_{\min}}, 1.11\cdots111 \times 2^{E_{\min}}]$)

  It is easy to check, using monotonicity of rounding, that:

  $$x \in [3\lambda/2, (2\lambda)^-] \implies \mathrm{fl}\langle 3\frac{x}{3}\rangle \in [3\lambda/2, ((2\lambda)^-)^-]$$

  From Table 3, we get for y a subnormal number:

  $$\mathrm{fl}(3y) = 1.1b_2' \cdots b_{p-1}' \times 2^{E_{\min}} \implies \mathrm{fl}(3y) = 3y$$

Hence:

$$\forall x \in [3\lambda/2, (2\lambda)^-] : \ \mathrm{fl}\langle 3\frac{x}{3}\rangle = 3(\frac{x}{3} + \eta_1) = x + 3\eta_1, |\eta_1| \leqslant \mu/2$$

From which we get:

$$|\mathrm{fl}\langle 3\frac{x}{3}\rangle - x| \leqslant \frac{3\mu}{2}$$

Consequently, we have:

$$\forall x \in [3\lambda/2, (2\lambda)^-] : \ \mathrm{fl}\langle 3\frac{x}{3}\rangle \in \{x^-, x, x^+\} \tag{20}$$

- If $x \in [2\lambda, (3\lambda)^-]$ (i.e., $x \in [1.0 \times 2^{E_{\min}+1}, 1.01\cdots111 \times 2^{E_{\min}+1}]$).

  It is trivial to show, using monotonicity of rounding, that:

  $$x \in [2\lambda, (3\lambda)^-] \implies \mathrm{fl}\langle 3\frac{x}{3}\rangle \in [2\lambda, ((3\lambda)^-)^-]$$

  In the interval $[2\lambda, ((3\lambda)^-)^-]$, the absolute distance between two consecutive floating-point numbers is 2μ (See Figure 2), which means that:

  $$\forall y : \ \mathrm{fl}\big(3y\big) \in [2\lambda, ((3\lambda)^-)^-] \implies |\mathrm{fl}\big(3y\big) - 3y| \leqslant \eta_3, |\eta_3| \leqslant \mu$$

  Therefore, for $x \in [2\lambda, (3\lambda)^-]$:

  $$\begin{aligned}\mathrm{fl}\langle 3\frac{x}{3}\rangle &= 3(\frac{x}{3} + \eta_1) + \eta_3, |\eta_1| \leqslant \mu/2, |\eta_3| \leqslant \mu \\ &= x + \eta, |\eta| \leqslant 5\mu/2\end{aligned}$$

  Then:

  $$|\mathrm{fl}\langle 3\frac{x}{3}\rangle - x| \leqslant \frac{5\mu}{2}$$

  The distance between consecutive floating-point numbers in the domain involved being 2μ, we eventually get:

  $$\forall x \in [2\lambda, (3\lambda)^-] : \ \mathrm{fl}\langle 3\frac{x}{3}\rangle \in \{x^-, x, x^+\} \tag{21}$$

  From Equations (19), (20), (21), we deduce that for $x \in [\lambda^+, (3\lambda)^-]$, $\mathrm{fl}\langle 3(x/3)\rangle$ is equal to x itself or one of its immediate predecessor or successor:

  $$\forall x \in [\lambda^+, (3\lambda)^-] : \ \mathrm{fl}\langle 3\frac{x}{3}\rangle \in \{x^-, x, x^+\} \tag{22}$$

5.3 Putting it all together

Knowing the worst case error for each operation, and armed with the tables from Sections 3 and 4, we may now determine the possible signs of the error when computing $\mathrm{fl}\langle(3x)/3\rangle$ and $\mathrm{fl}\langle 3(x/3)\rangle$. Consider for example the floating-point number x of the form $1.1b_2 \ldots b_{p-4}001 \times 2^{e_x}$ and the expression $(3x)/3$. Using Table 2, we get all the possible forms for the significand of $y = \mathrm{fl}(3x)$. They are—in the absence of underflow:

$$\begin{cases} 1.0b'_2 \ldots b'_{p-4}001 \\ 1.0b'_2 \ldots b'_{p-4}011 \\ 1.0b'_2 \ldots b'_{p-4}101 \\ 1.0b'_2 \ldots b'_{p-4}111 \end{cases} \tag{23}$$

Table 10: Result and sign of the error when computing $\mathrm{fl}\langle(3x)/3\rangle$ without underflow.

$\mathbf{x}$	$\mathrm{fl}\langle(\mathbf{3x})/\mathbf{3}\rangle$	Error
$1.0b_2\cdots b_{p-4}000$	$1.0b'_2\cdots b'_{p-4}000$	EQ
$1.0b_2\cdots b_{p-4}001$	$1.0b'_2\cdots b'_{p-4}001$	EQ
$1.0b_2\cdots b_{p-4}010$	$1.0b'_2\cdots b'_{p-4}011$	GT
	$1.0b'_2\cdots b'_{p-4}010$	EQ
$1.0b_2\cdots b_{p-4}011$	$1.0b'_2\cdots b'_{p-4}011$	EQ
$1.0b_2\cdots b_{p-4}100$	$1.0b'_2\cdots b'_{p-4}100$	EQ
$1.0b_2\cdots b_{p-4}101$	$1.0b'_2\cdots b'_{p-4}101$	EQ
$1.0b_2\cdots b_{p-4}110$	$1.0b'_2\cdots b'_{p-4}110$	EQ
	$1.0b'_2\cdots b'_{p-4}101$	LT
$1.0b_2\cdots b_{p-4}111$	$1.0b'_2\cdots b'_{p-4}111$	EQ
$1.1b_2\cdots b_{p-4}000$	$1.1b'_2\cdots b'_{p-4}000$	EQ
$1.1b_2\cdots b_{p-4}001$	$1.1b'_2\cdots b'_{p-4}001$	EQ
$1.1b_2\cdots b_{p-4}010$	$1.1b'_2\cdots b'_{p-4}011$	GT
$1.1b_2\cdots b_{p-4}011$	$1.1b'_2\cdots b'_{p-4}011$	EQ
$1.1b_2\cdots b_{p-4}100$	$1.1b'_2\cdots b'_{p-4}100$	EQ
$1.1b_2\cdots b_{p-4}101$	$1.1b'_2\cdots b'_{p-4}101$	EQ
$1.1b_2\cdots b_{p-4}110$	$1.1b'_2\cdots b'_{p-4}101$	LT
$1.1b_2\cdots b_{p-4}111$	$1.1b'_2\cdots b'_{p-4}111$	EQ

For each of these four forms, we get from Table 4 all the possible forms for the significand of $\mathrm{fl}\big(y/3\big)$. For example, for the first form $1.0b'_2\ldots b'_{p-4}001$, we get:

$$\begin{cases} 1.0b''_2\ldots b''_{p-4}001 \\ 1.1b''_2\ldots b''_{p-4}001 \\ 1.0b''_2\ldots b''_{p-4}100 \\ 1.1b''_2\ldots b''_{p-4}100 \\ 1.0b''_2\ldots b''_{p-4}111 \\ 1.1b''_2\ldots b''_{p-4}111 \end{cases} \tag{24}$$

However, from Eq. (15), we know that $\mathrm{fl}\langle(3x)/3\rangle$ for $x = 1.1b_2\ldots b_{p-4}001\times 2^{e_x}$ must be one of the following floating-point numbers:

$$\begin{cases} 1.1b_2\ldots b_{p-4}000\times 2^{e_x} \\ 1.1b_2\ldots b_{p-4}001\times 2^{e_x} \\ 1.1b_2\ldots b_{p-4}010\times 2^{e_x} \end{cases}$$

We can then discard from Eq. (24) all forms but the second one: $1.1b_2\ldots b_{p-4}001$. If we do the same work for all forms in Eq. (23), we will find that the only possible significand for $\mathrm{fl}\langle(3x)/3\rangle$ is $1.1b_2\ldots b_{p-4}001$, that is the one of x itself.

Doing the work exemplified above for all possible significands for x that do not lead to an underflow in either the division or the multiplication, we get Table 10. As proved in Section 5.2.1, should an underflow arise during the computation of $\mathrm{fl}\langle(3x)/3\rangle$, it is not necessary to compute such a table since we then always have $\mathrm{fl}\langle(3x)/3\rangle = x$.

Computing the same table for the expression $\mathrm{fl}\langle 3(x/3)\rangle$ results in a table with 96—that is, $2^5\times 3$—entries: given $x = b_0.b_1b_2\cdots b_{p-3}b_{p-2}b_{p-1}\times 2^{e_x}$, $\mathrm{fl}\langle 3(x/3)\rangle$ may always be x or one of its two immediate neighbours, whatever the values of b_0, b_1, b_{p-3}, b_{p-2}, and b_{p-1}. Therefore, it is useless to display the resulting table as it is uninformative.

6 Conclusion

Using both traditional error analysis and the tables computed in Sections 3 and 4, we have been able to compute refined error bounds in Section 5 for $\mathrm{fl}\langle 3x/3\rangle$ and $\mathrm{fl}\langle 3(x/3)\rangle$. At first sight (See Tables 11 and 12), it seems that the expression $3x/3$ gives overall more accurate results than the expression $3(x/3)$ whenever an underflow occurs, and that no expression is better than the other when no underflow occurs.

Table 11: Refined worst case error analysis for $\mathrm{fl}\langle\frac{3x}{3}\rangle$, with $|\delta| < 2u$.

$y = 3x$ \ $z = y/3$	Underflow	No underflow
Underflow	x	—
No underflow	x	$x(1+\delta)$

Table 12: Refined worst case error analysis for $\mathrm{fl}\langle 3\frac{x}{3}\rangle$, with: $|\eta| \leqslant \mu$, $|\delta| < 2u$.

$y = x/3$ \ $z = 3y$	Underflow	No underflow
Underflow	$\{x^-, x, x^+\}$	$\{x^-, x, x^+\}$
No underflow	—	$x(1+\delta)$

This is only part of the story, however. Our detailed analysis reveals that only when the last three bits of the significand of x are "010" or "110" may the expression $3x/3$ be rounded to a value different from x (see Table 10), while the same situation may occur whatever the value of x's significand with the expression $3(x/3)$ (refer to the end of the previous section). The expression $3x/3$ is therefore to be prefered to $3(x/3)$ in all cases.

Tables 1, 2, 3, 4, 5, 6, and 7 are all interesting and useful in their own right whenever one needs to obtain more information on the division and multiplication by three of a floating-point number than merely a worst-case error bound. We expect to reuse the method used here to prove properties regarding the expressions $3x/3$ and $3(x/3)$, which combines traditional error analysis with systematic study of the behaviour of equivalence classes of floating-point numbers, in proving the correctness of polyadic splitting operators for interval analysis [5].

The question of multiplying and dividing a number by another in floating-point arithmetic has been addressed several times already in slightly different settings: Kahan [3, thm. 7] showed that $\mathrm{fl}\langle(m/n) \times n\rangle = m$ for integers m and n such that $|m| < 2^{p-1}$ and n is the sum of two powers of 2. Edelman [2] investigated the problem of finding the smallest positive floating-point x such that $x(1/x) \neq 1$.

Obviously, the next step for us is to address the problem presented in this paper for other integer values than 3. That might prove indeed necessary for our work on polyadic splitting operators. Once the preliminary worst case error analysis of an expression is performed manually, a large part of the proofs can be performed programmatically, which departs from the more involved—though more powerful and flexible—approaches used by, e.g., Kahan and Edelman. Generalizing the approach to other similar problems of interest is a direction of future research.

References

[1] Florent De Dinechin and Laurent-Stéphane Didier. "Table-Based Division by Small Integer Constants". In: *Reconfigurable Computing: Architectures, Tools and Applications*. Ed. by OliverC. S. Choy et al. Vol. 7199. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, pp. 53–63. ISBN: 978-3-642-28364-2. DOI: 10.1007/978-3-642-28365-9_5.

[2] Alan Edelman. "When is $x \times (1/x) \neq 1$?" Unpublished note. Dec. 1994.

[3] David Goldberg. "What every computer scientist should know about floating-point arithmetic". In: *ACM Computing Surveys* 23.1 (Mar. 1991), pp. 5–48. ISSN: 0360-0300. DOI: 10.1145/103162.103163.

[4] Frédéric Goualard. "How do you compute the midpoint of an interval?" In: *ACM Transactions on Mathematical Software* 40.2 (Feb. 2014). DOI: 10.1145/2493882.

[5] Frédéric Goualard and Laurent Granvilliers. "Polyadic Splitting of FLoating-point Intervals". In preparation. 2017.

[6] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Second edition. Society for Industrial and Applied Mathematics, 2002. ISBN: 0898715210.

[7] *IEEE Standard for Binary Floating-Point Arithmetic*. American National Standard (ANSI) IEEE Std 754-1985. The Institute of Electrical and Electronics Engineers, Inc, 1985.

[8] Claude-Pierre Jeannerod and Siegfried M. Rump. "On relative errors of floating-point operations: optimal bounds and applications". In: *Mathematics of Computation* (2017). To appear.

[9] Jean-Michel Muller et al. *Handbook of Floating-Point Arithmetic*. Birkhauser Boston Inc, Dec. 2008. ISBN: 978-0817647049.

A Multiplication by 3 involving no denormal number

The two columns below present the results of the multiplication by 3 of a positive floating-point number x depending on the last three bits of its significand. Thanks to Lemmas 1 and 2, we know that we also need to consider the first bit after the radix point to know how many bits from the result to discard. The left column (resp. right column) corresponds to the cases for which we need to discard the last bit (resp. the last two bits) underlined. For convenience, we display only the significands and leave the exponents implicit.

		1.	0	b_2	$\cdots$	b_{p-4}	0	0	0
$\times$		1.	1						
		1.	0	b_2	$\cdots$	b_{p-4}	0	0	0
$+$	1.	0	b_2	$\cdots$	b_{p-4}	0	0	0	
	1.	1	b'_2	$\cdots$	b'_{p-4}	b'_{p-3}	0	0	$\underline{0}$

		1.	b_1	b_2	$\cdots$	b_{p-4}	0	0	0
$\times$		1.	1						
		1.	b_1	b_2	$\cdots$	b_{p-4}	0	0	0
$+$	1.	b_1	b_2	$\cdots$	b_{p-4}	0	0	0	
1.	0	b'_2		$\cdots$		b'_{p-3}	b'_{p-2}	0	$\underline{0}$	$\underline{0}$

		1.	0	b_2	$\cdots$	b_{p-4}	0	0	1
$\times$		1.	1						
		1.	0	b_2	$\cdots$	b_{p-4}	0	0	1
$+$	1.	0	b_2	$\cdots$	b_{p-4}	0	0	1	
	1.	1	b'_2	$\cdots$	b'_{p-4}	b'_{p-3}	0	1	$\underline{1}$

		1.	b_1	b_2	$\cdots$	b_{p-4}	0	0	1
$\times$		1.	1						
		1.	b_1	b_2	$\cdots$	b_{p-4}	0	0	1
$+$	1.	b_1	b_2	$\cdots$	b_{p-4}	0	0	1	
1.	0	b'_2		$\cdots$		b'_{p-3}	b'_{p-2}	0	$\underline{1}$	$\underline{1}$

		1.	0	b_2	$\cdots$	b_{p-4}	0	1	0
$\times$		1.	1						
		1.	0	b_2	$\cdots$	b_{p-4}	0	1	0
$+$	1.	0	b_2	$\cdots$	b_{p-4}	0	1	0	
	1.	1	b'_2	$\cdots$	b'_{p-4}	b'_{p-3}	1	1	$\underline{0}$

		1.	b_1	b_2	$\cdots$	b_{p-4}	0	1	0
$\times$		1.	1						
		1.	b_1	b_2	$\cdots$	b_{p-4}	0	1	0
$+$	1.	b_1	b_2	$\cdots$	b_{p-4}	0	1	0	
1.	0	b'_2		$\cdots$		b'_{p-3}	b'_{p-2}	1	$\underline{1}$	$\underline{0}$

		1.	0	b_2	$\cdots$	b_{p-4}	0	1	1
$\times$		1.	1						
		1.	0	b_2	$\cdots$	b_{p-4}	0	1	1
$+$	1.	0	b_2	$\cdots$	b_{p-4}	0	1	1	
	1.	1	b'_2	$\cdots$	b'_{p-4}	b'_{p-3}	0	0	$\underline{1}$

		1.	b_1	b_2	$\cdots$	b_{p-4}	0	1	1
$\times$		1.	1						
		1.	b_1	b_2	$\cdots$	b_{p-4}	0	1	1
$+$	1.	b_1	b_2	$\cdots$	b_{p-4}	0	1	1	
1.	0	b'_2		$\cdots$		b'_{p-3}	b'_{p-2}	0	$\underline{0}$	$\underline{1}$

		1.	0	b_2	$\cdots$	b_{p-4}	1	0	0
$\times$		1.	1						
		1.	0	b_2	$\cdots$	b_{p-4}	1	0	0
$+$	1.	0	b_2	$\cdots$	b_{p-4}	1	0	0	
	1.	1	b'_2	$\cdots$	b'_{p-4}	b'_{p-3}	1	0	$\underline{0}$

		1.	b_1	b_2	$\cdots$	b_{p-4}	1	0	0
$\times$		1.	1						
		1.	b_1	b_2	$\cdots$	b_{p-4}	1	0	0
$+$	1.	b_1	b_2	$\cdots$	b_{p-4}	1	0	0	
1.	0	b'_2		$\cdots$		b'_{p-3}	b'_{p-2}	1	$\underline{0}$	$\underline{0}$

		1.	0	b_2	$\cdots$	b_{p-4}	1	0	1
$\times$		1.	1						
		1.	0	b_2	$\cdots$	b_{p-4}	1	0	1
$+$	1.	0	b_2	$\cdots$	b_{p-4}	1	0	1	
	1.	1	b'_2	$\cdots$	b'_{p-4}	b'_{p-3}	1	1	$\underline{1}$

		1.	b_1	b_2	$\cdots$	b_{p-4}	1	0	1
$\times$		1.	1						
		1.	b_1	b_2	$\cdots$	b_{p-4}	1	0	1
$+$	1.	b_1	b_2	$\cdots$	b_{p-4}	1	0	1	
1.	0	b'_2		$\cdots$		b'_{p-3}	b'_{p-2}	1	$\underline{1}$	$\underline{1}$

		1.	0	b_2	$\cdots$	b_{p-4}	1	1	0
$\times$		1.	1						
		1.	0	b_2	$\cdots$	b_{p-4}	1	1	0
$+$	1.	0	b_2	$\cdots$	b_{p-4}	1	1	0	
	1.	1	b'_2	$\cdots$	b'_{p-4}	b'_{p-3}	0	1	$\underline{0}$

		1.	b_1	b_2	$\cdots$	b_{p-4}	1	1	0
$\times$		1.	1						
		1.	b_1	b_2	$\cdots$	b_{p-4}	1	1	0
$+$	1.	b_1	b_2	$\cdots$	b_{p-4}	1	1	0	
1.	0	b'_2		$\cdots$		b'_{p-3}	b'_{p-2}	0	$\underline{1}$	$\underline{0}$

		1.	0	b_2	$\cdots$	b_{p-4}	1	1	1
×		1.	1						
		1.	0	b_2	$\cdots$	b_{p-4}	1	1	1
+	1.	0	b_2	$\cdots$	b_{p-4}	1	1	1	
	1.	1	b'_2	$\cdots$	b'_{p-4}	b'_{p-3}	1	0	$\underline{1}$

		1.	b_1	b_2	$\cdots$	b_{p-4}	1	1	1
×		1.	1						
		1.	b_1	b_2	$\cdots$	b_{p-4}	1	1	1
+	1.	b_1	b_2	$\cdots$	b_{p-4}	1	1	1	
1.	0	b'_2		$\cdots$		b'_{p-3}	b'_{p-2}	1	$\underline{0}$	$\underline{1}$

B Multiplication by 3 involving denormal numbers

List of the two different results for all possible values of the rightmost bits, on the left for a p bits significand, and on the right for a $p+1$ bits significand, where the bit to discard is underlined.

		0.	1	b_2	$\cdots$	b_{p-4}	0	0	0
×		1.	1						
		0.	1	b_2	$\cdots$	b_{p-4}	0	0	0
+	0.	1	b_2	$\cdots$	b_{p-4}	0	0	0	
	0.	1	b'_2	$\cdots$	b'_{p-4}	b'_{p-3}	0	0	0

		0.	1	b_2	$\cdots$	b_{p-4}	0	0	0
×		1.	1						
		0.	1	b_2	$\cdots$	b_{p-4}	0	0	0
+	0.	1	b_2	$\cdots$	b_{p-4}	0	0	0	
1.	0	b'_2		$\cdots$		b'_{p-2}	0	0	$\underline{0}$

		0.	1	b_2	$\cdots$	b_{p-4}	0	0	1
×		1.	1						
		0.	1	b_2	$\cdots$	b_{p-4}	0	0	1
+	0.	1	b_2	$\cdots$	b_{p-4}	0	0	1	
	0.	1	b'_2	$\cdots$	b'_{p-4}	b'_{p-3}	0	1	1

		0.	1	b_2	$\cdots$	b_{p-4}	0	0	1
×		1.	1						
		0.	1	b_2	$\cdots$	b_{p-4}	0	0	1
+	0.	1	b_2	$\cdots$	b_{p-4}	0	0	1	
1.	0	b'_2		$\cdots$		b'_{p-2}	0	1	$\underline{1}$

		0.	1	b_2	$\cdots$	b_{p-4}	0	1	0
×		1.	1						
		0.	1	b_2	$\cdots$	b_{p-4}	0	1	0
+	0.	1	b_2	$\cdots$	b_{p-4}	0	1	0	
	0.	1	b'_2	$\cdots$	b'_{p-4}	b'_{p-3}	1	1	0

		0.	1	b_2	$\cdots$	b_{p-4}	0	1	0
×		1.	1						
		0.	1	b_2	$\cdots$	b_{p-4}	0	1	0
+	0.	1	b_2	$\cdots$	b_{p-4}	0	1	0	
1.	0	b'_2		$\cdots$		b'_{p-2}	1	1	$\underline{0}$

		0.	1	b_2	$\cdots$	b_{p-4}	0	1	1
×		1.	1						
		0.	1	b_2	$\cdots$	b_{p-4}	0	1	1
+	0.	1	b_2	$\cdots$	b_{p-4}	0	1	1	
	0.	1	b'_2	$\cdots$	b'_{p-4}	b'_{p-3}	0	0	1

		0.	1	b_2	$\cdots$	b_{p-4}	0	1	1
×		1.	1						
		0.	1	b_2	$\cdots$	b_{p-4}	0	1	1
+	0.	1	b_2	$\cdots$	b_{p-4}	0	1	1	
1.	0	b'_2		$\cdots$		b'_{p-2}	0	0	$\underline{1}$

		0.	1	b_2	$\cdots$	b_{p-4}	1	0	0
×		1.	1						
		0.	1	b_2	$\cdots$	b_{p-4}	1	0	0
+	0.	1	b_2	$\cdots$	b_{p-4}	1	0	0	
	0.	1	b'_2	$\cdots$	b'_{p-4}	b'_{p-3}	1	0	0

		0.	1	b_2	$\cdots$	b_{p-4}	1	0	0
×		1.	1						
		0.	1	b_2	$\cdots$	b_{p-4}	1	0	0
+	0.	1	b_2	$\cdots$	b_{p-4}	1	0	0	
1.	0	b'_2	$\cdots$	b'_{p-4}	b'_{p-2}	1	0	$\underline{0}$	

		0.	1	b_2	$\cdots$	b_{p-4}	1	0	1
×		1.	1						
		0.	1	b_2	$\cdots$	b_{p-4}	1	0	1
+	0.	1	b_2	$\cdots$	b_{p-4}	1	0	1	
	0.	1	b'_2		$\cdots$	b'_{p-3}	1	1	1

		0.	1	b_2	$\cdots$	b_{p-4}	1	0	1
×		1.	1						
		0.	1	b_2	$\cdots$	b_{p-4}	1	0	1
+	0.	1	b_2	$\cdots$	b_{p-4}	1	0	1	
1.	0	b'_2		$\cdots$		b'_{p-2}	1	1	$\underline{1}$

$$
\begin{array}{cccccccccc}
 & & 0. & 1 & b_2 & \cdots & b_{p-4} & 1 & 1 & 0 \\
\times & & 1. & 1 & & & & & & \\
\hline
 & & 0. & 1 & b_2 & \cdots & b_{p-4} & 1 & 1 & 0 \\
+ & 0. & 1 & b_2 & \cdots & b_{p-4} & 1 & 1 & 0 & \\
\hline
 & 0. & 1 & b'_2 & \cdots & b'_{p-4} & b'_{p-3} & 0 & 1 & 0
\end{array}
\qquad
\begin{array}{cccccccccc}
 & & 0. & 1 & b_2 & \cdots & b_{p-4} & 1 & 1 & 0 \\
\times & & 1. & 1 & & & & & & \\
\hline
 & & 0. & 1 & b_2 & \cdots & b_{p-4} & 1 & 1 & 0 \\
+ & 0. & 1 & b_2 & \cdots & b_{p-4} & 1 & 1 & 0 & \\
\hline
 & 1. & 0 & b'_2 & & \cdots & b'_{p-2} & 0 & 1 & \underline{0}
\end{array}
$$

$$
\begin{array}{cccccccccc}
 & & 0. & 1 & b_2 & \cdots & b_{p-4} & 1 & 1 & 1 \\
\times & & 1. & 1 & & & & & & \\
\hline
 & & 0. & 1 & b_2 & \cdots & b_{p-4} & 1 & 1 & 1 \\
+ & 0. & 1 & b_2 & \cdots & b_{p-4} & 1 & 1 & 1 & \\
\hline
 & 0. & 1 & b'_2 & \cdots & b'_{p-4} & b'_{p-3} & 1 & 0 & 1
\end{array}
\qquad
\begin{array}{cccccccccc}
 & & 0. & 1 & b_2 & \cdots & b_{p-4} & 1 & 1 & 1 \\
\times & & 1. & 1 & & & & & & \\
\hline
 & & 0. & 1 & b_2 & \cdots & b_{p-4} & 1 & 1 & 1 \\
+ & 0. & 1 & b_2 & \cdots & b_{p-4} & 1 & 1 & 1 & \\
\hline
 & 1. & 0 & b'_2 & & \cdots & b'_{p-2} & 1 & 0 & \underline{1}
\end{array}
$$

C Division by 3 involving no denormal number

The leftmost column considers $x = 1.0b_2 \cdots b_{p-1} \times 2^{e_x}$ and the rightmost column considers $x = 1.1b_2 \cdots b_{p-1} \times 2^{e_x}$. For each triplet of possible last three bits of the significand of x, we consider the three possible remainders on b_{p-4}.

When x ends in "000"

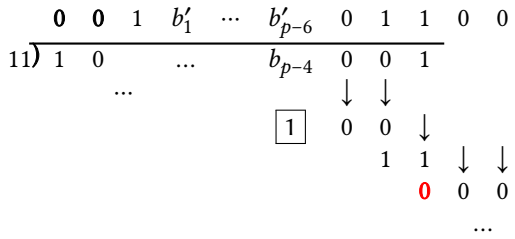
$$
\begin{array}{rcccccccccc}
 & \mathbf{0} & \mathbf{0} & 1 & b'_1 & \cdots & b'_{p-6} & 0 & 0 & 0 & 0 & 0 \\
11) & 1 & 0 & & \ldots & & b_{p-4} & 0 & 0 & 0 & & \\
 & & \ldots & & & & & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
 & & & & & & \boxed{0} & 0 & 0 & \mathbf{0} & 0 & 0
\end{array}
\qquad
\begin{array}{rcccccccccc}
 & \mathbf{0} & 1 & 0 & b'_2 & \cdots & b'_{p-5} & 0 & 0 & 0 & 0 \\
11) & 1 & 1 & & \ldots & & b_{p-4} & 0 & 0 & 0 & \\
 & & \ldots & & & & & \downarrow & \downarrow & \downarrow & \downarrow \\
 & & & & & & \boxed{0} & 0 & 0 & \mathbf{0} & 0
\end{array}
$$

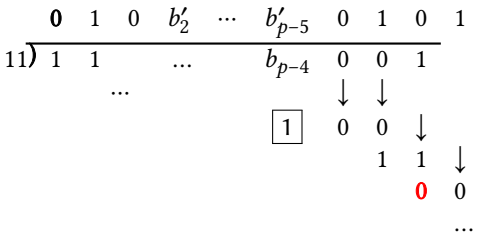
$$
\begin{array}{rcccccccccc}
 & \mathbf{0} & \mathbf{0} & 1 & b'_1 & \cdots & b'_{p-6} & 0 & 1 & 0 & 1 & 0 \\
11) & 1 & 0 & & \ldots & & b_{p-4} & 0 & 0 & 0 & & \\
 & & \ldots & & & & & \downarrow & \downarrow & & & \\
 & & & & & & \boxed{1} & 0 & 0 & \downarrow & \downarrow & \\
 & & & & & & & & \mathbf{1} & \mathbf{0} & 0 & \downarrow \\
 & & & & & & & & & & 1 & 0 \\
 & & & & & & & & & & \ldots &
\end{array}
\qquad
\begin{array}{rcccccccccc}
 & \mathbf{0} & 1 & 0 & b'_2 & \cdots & b'_{p-5} & 0 & 1 & 0 & \mathbf{1} \\
11) & 1 & 1 & & \ldots & & b_{p-4} & 0 & 0 & 0 & \\
 & & \ldots & & & & & \downarrow & \downarrow & & \\
 & & & & & & \boxed{1} & 0 & 0 & \downarrow & \downarrow \\
 & & & & & & & & \mathbf{1} & \mathbf{0} & 0 \\
 & & & & & & & & & & 1 \\
 & & & & & & & & & & \ldots
\end{array}
$$

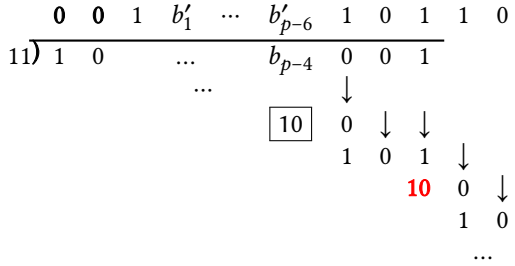
$$
\begin{array}{rcccccccccc}
 & \mathbf{0} & \mathbf{0} & 1 & b'_1 & \cdots & b'_{p-6} & 1 & 0 & 1 & 0 & 1 \\
11) & 1 & 0 & & \ldots & & b_{p-4} & 0 & 0 & 0 & & \\
 & & & & \ldots & & & \downarrow & & & & \\
 & & & & & & \boxed{10} & 0 & \downarrow & \downarrow & & \\
 & & & & & & & 1 & 0 & 0 & \downarrow & \downarrow \\
 & & & & & & & & & \mathbf{1} & 0 & 0 \\
 & & & & & & & & & & \ldots &
\end{array}
\qquad
\begin{array}{rcccccccccc}
 & \mathbf{0} & 1 & 0 & b'_2 & \cdots & b'_{p-5} & 1 & 0 & 1 & 0 \\
11) & 1 & 1 & & \ldots & & b_{p-4} & 0 & 0 & 0 & \\
 & & \ldots & & & & & \downarrow & & & \\
 & & & & & & \boxed{10} & 0 & \downarrow & \downarrow & \\
 & & & & & & & 1 & 0 & 0 & \downarrow \\
 & & & & & & & & & \mathbf{1} & 0 \\
 & & & & & & & & & & \ldots
\end{array}
$$

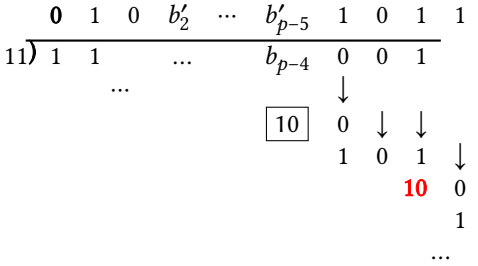
When x ends in "001"

$$
\begin{array}{rcccccccccc}
 & \mathbf{0} & \mathbf{0} & 1 & b'_1 & \cdots & b'_{p-6} & 0 & 0 & 0 & 0 & 1 \\
11) & 1 & 0 & & \ldots & & b_{p-4} & 0 & 0 & 1 & & \\
 & & \ldots & & & & & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
 & & & & & & \boxed{0} & 0 & 0 & \mathbf{1} & 0 & 0 \\
 & & & & & & & & & & & 1 \\
 & & & & & & & & & & & \ldots
\end{array}
\qquad
\begin{array}{rcccccccccc}
 & \mathbf{0} & 1 & 0 & b'_2 & \cdots & b'_{p-5} & 0 & 0 & 0 & 0 \\
11) & 1 & 1 & & \ldots & & b_{p-4} & 0 & 0 & 1 & \\
 & & \ldots & & & & & \downarrow & \downarrow & \downarrow & \downarrow \\
 & & & & & & \boxed{0} & 0 & 0 & \mathbf{1} & 0 \\
 & & & & & & & & & & \ldots
\end{array}
$$

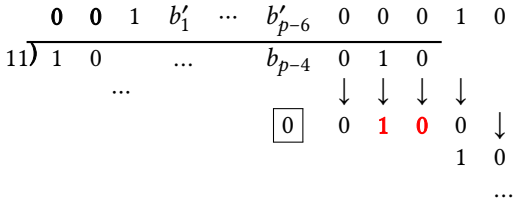


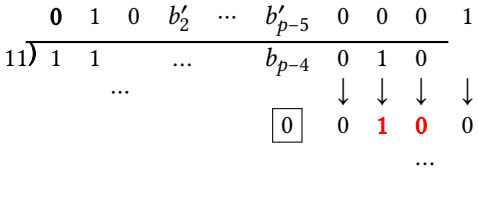


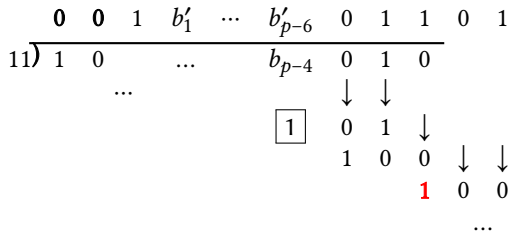


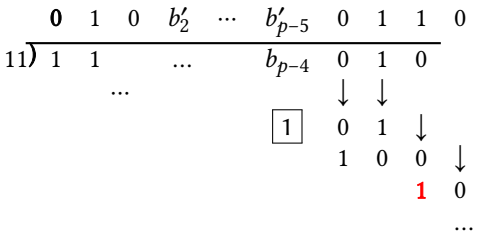


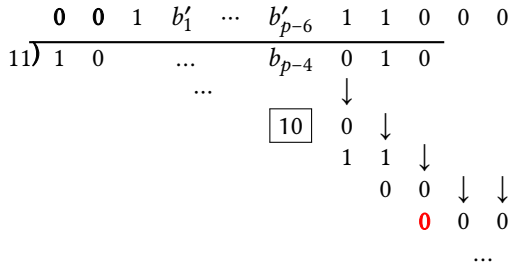
When x ends in "010"

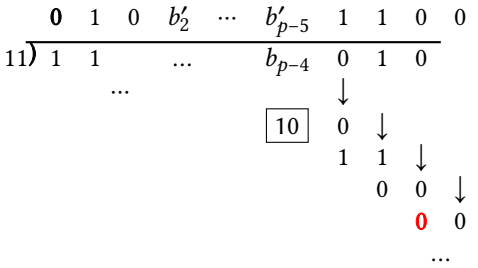




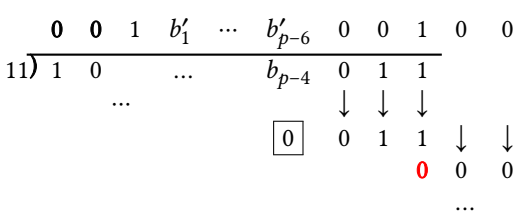


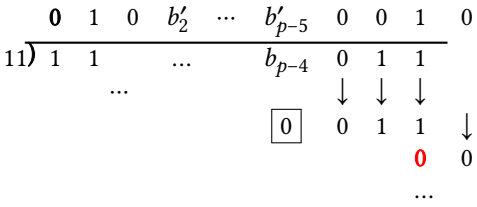


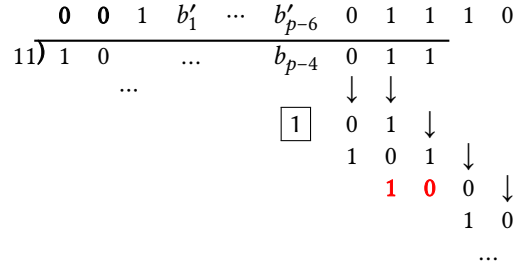


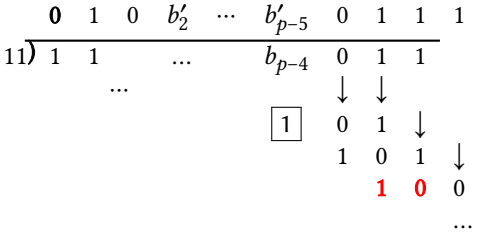


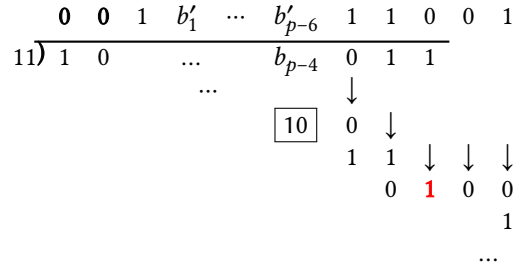
When x ends in "011"

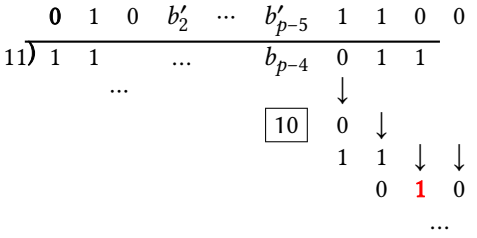




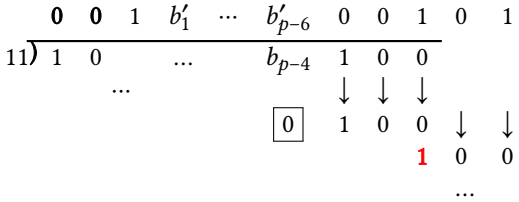


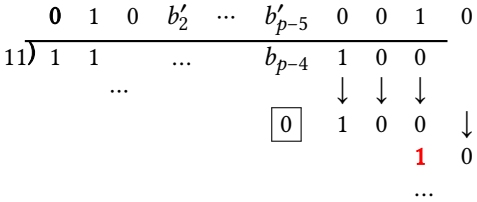


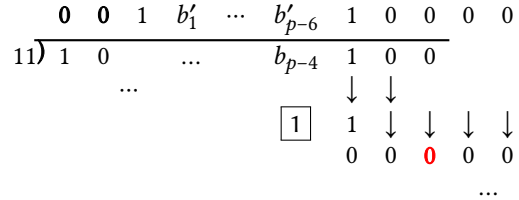


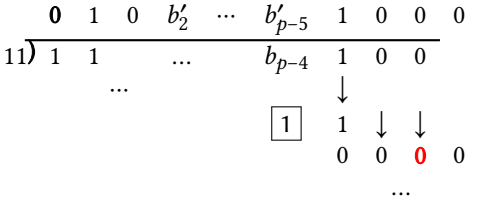


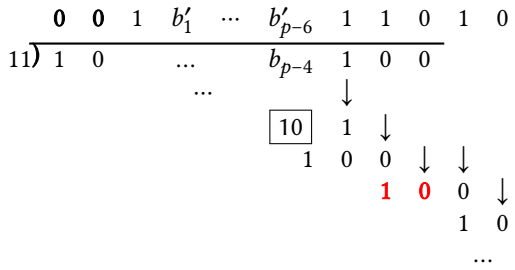
When x ends in "100"

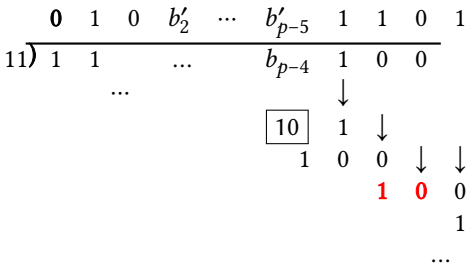




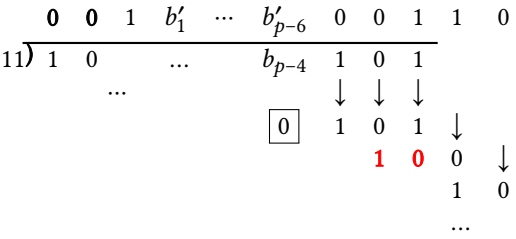


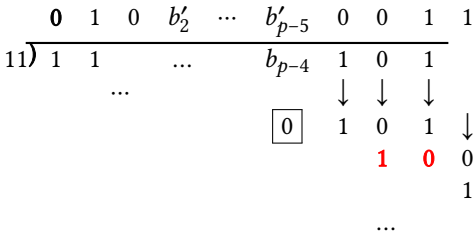


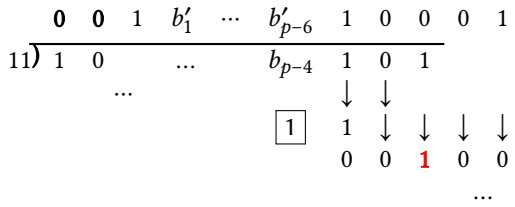


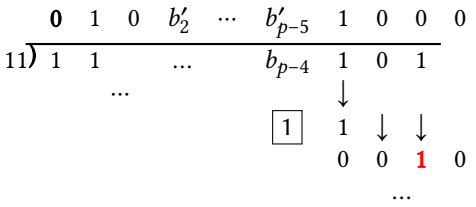


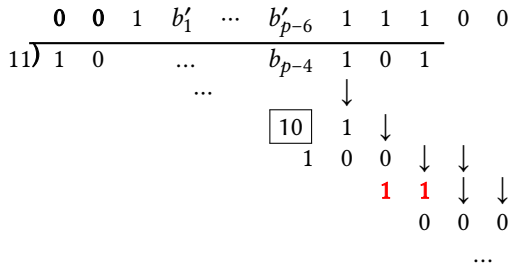
When x ends in "101"

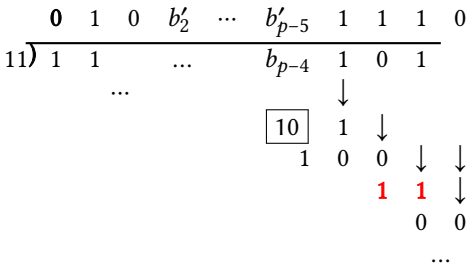




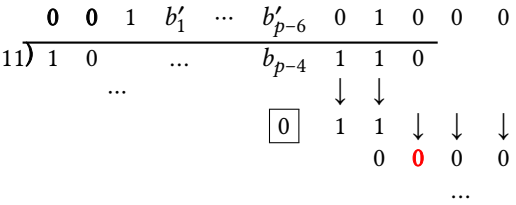


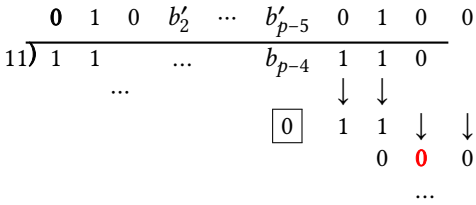


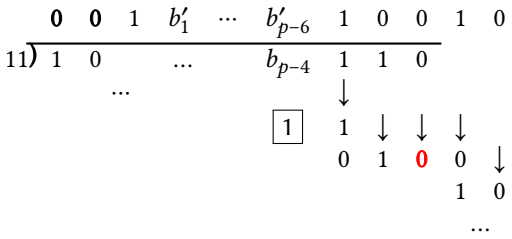


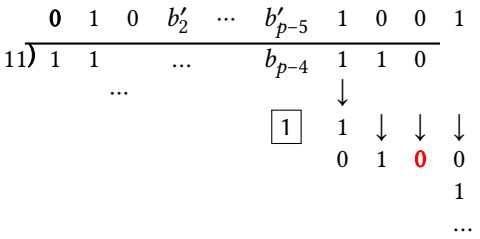


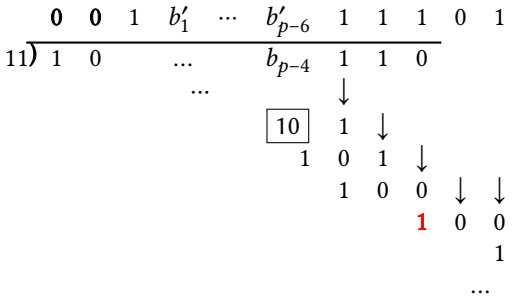
When x ends in "110"

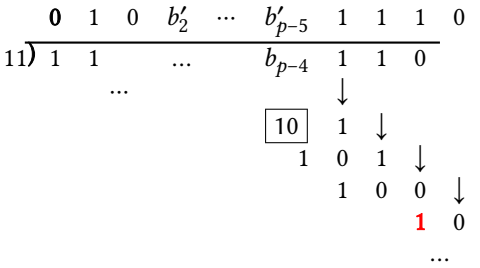




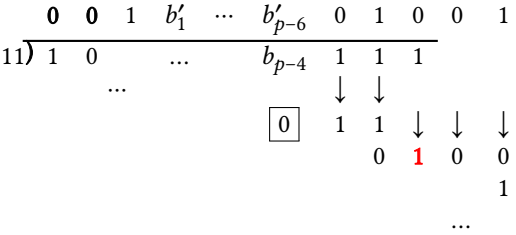


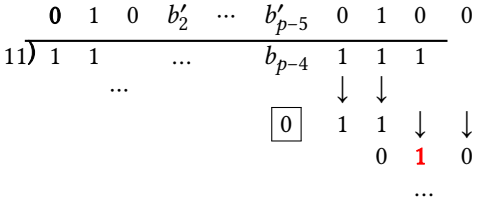


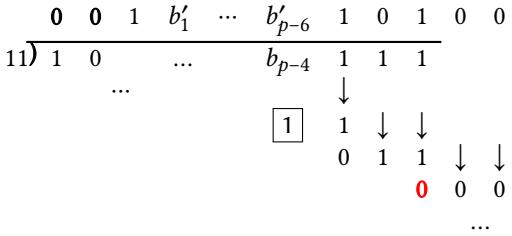


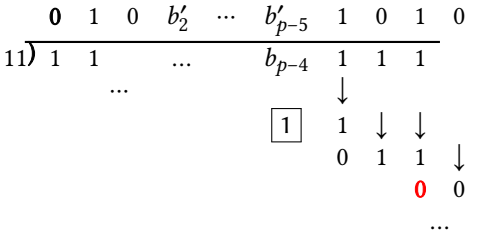


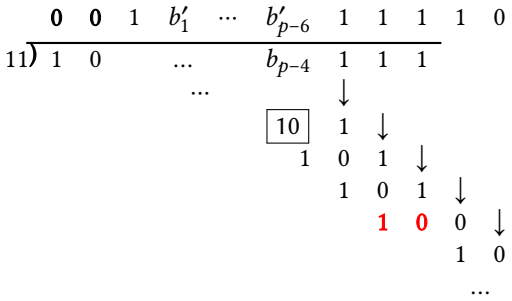
When x ends in "111"

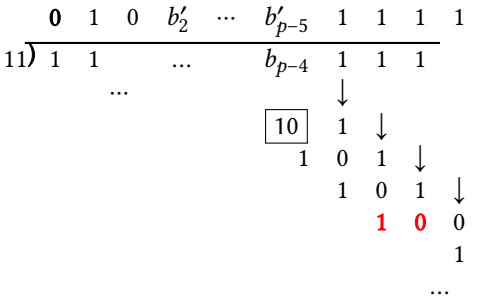












D Division by 3 involving denormal numbers

.

As explained in Section 4.2, the rules for the division involving denormal numbers may be inferred from the ones for the division involving normal numbers (Refer to Section C above).

Rapport de recherche

Research Report

When $(3x)/3$ and $3(x/3)$ are not equal to x

RR n° xx.xx

Frédéric Goualard

LS2N
LABORATOIRE
DES SCIENCES
DU NUMÉRIQUE
DE NANTES