On the definition of $\operatorname{ulp}(x)$

Jean-Michel Muller
CNRS – Laboratoire LIP, projet Arenaire
(CNRS, ENS Lyon, INRIA, Univ. Lyon 1),
46 Allée d'Italie,
69364 Lyon Cedex 07,
France

Function ulp (acronym for *unit in the last place*) is frequently used for expressing errors in floating-point computations. We present several previously suggested definitions of that function, and analyse some of their properties.

Categories and Subject Descriptors: G.1.0 [**Numerical Analysis**]: General—*Computer arithmetic*

Additional Key Words and Phrases: Floating-point arithmetic, computer arithmetic, unit in the last place, ULP

1. INTRODUCTION

The term ulp (acronym for *unit in the last place*) was coined by W. Kahan in 1960. The original definition was [Kahan 2004]:

> $\operatorname{ulp}(x)$ is the gap between the two floating-point numbers nearest x, even if x is one of them.

Ulps are of interest for measuring/describing the accuracy of "atomic" computations (e.g., elementary functions, complex division, evaluation of small polynomials or dot products), that is, computations that need be extremely accurate. In a way, ulps are interesting for expressing errors when these errors are not larger than a few ulps. They are not suited for expressing errors of fairly "large" computations. A consequence of that is that the use of ulps is widespread in computer arithmetic and quite infrequent in numerical analysis.

As told by Kahan [Kahan 2004], the adoption of the IEEE-754 standard for floating-point arithmetic has made infinities and NaNs ubiquitous, and that must be taken into account in the definition of $\operatorname{ulp}(x)$. Kahan now suggests the following definition:

> $\operatorname{ulp}(x)$ is the gap between the two *finite* floating-point numbers nearest

Authors' address: J.-M. Muller, LIP/Arénaire (CNRS-ENS Lyon-INRIA-UCBL), ENS Lyon, 46 Allée d'Italie, F-69364 Lyon Cedex 07 France; email: Jean-Michel.Muller@ens-lyon.fr.

x, even if x is one of them. (But $\mathrm{ulp}\,(\mathrm{NaN})$ is NaN.)

Several slightly different definitions of $\mathrm{ulp}\,(x)$ appear in the literature [Goldberg 1991; Harrison 1999; Markstein 2000; Overton 2001]. In this paper, we restate these various definitions and we analyze some of their properties. Among these properties, some have certainly already been found by other people having dealt with this topic (without, to my knowledge, having been published, except when I give references). And yet, I feel it may be useful to collect them in a paper. Good knowledge of these properties may be important, for instance, for anyone who want to prove sure yet tight bounds on the errors of atomic computations: more than claiming that a definition is better than the other ones, the goal of the paper is to explain which properties are true, in which cases, so that they can be safely used in proofs.

We briefly define the floating-point numbers, to make the paper self-complete. For more information, see [Kahan 1996; Overton 2001].

Definition 1 FLOATING-POINT NUMBERS. *A floating-point format is (at least partially*[1]*) defined by three integers:* a radix $r \geq 2$, *a* precision p *and extremal exponents* $E_{min} < 0$ *and* $E_{max} > 0$.

In such a format, a finite floating-point number X *has the form*

$$X = (-1)^s \times M \times r^{e-p+1},$$

where s *is* 0 *or* 1*, and* M *and* e *are integers satisfying* $0 \leq M \leq r^p - 1$ *and* $E_{min} \leq e \leq E_{max}$. *Whenever possible, it is required that* $r^{p-1} \leq M$. *In such a case,* X *is a* normal *number. When this is not possible (that is, when* $|X| < r^{Emin}$*),* X *is a* subnormal *number.*

The real number $m = M \times r^{-p+1}$ *is called the* mantissa *of* X*, and* e *is called the* exponent *of* X*. The* precision p *is the number of radix-*r *digits of the mantissa*[2]*.*

Throughout the paper, we assume a radix-r floating-point (FP for short) arithmetic, with precision p. If X is an FP number, then X^+ denotes the smallest FP number larger than X and X^- denotes the largest FP number less than X.

A good definition of function ulp:

—should (of course) agree with the "intuitive" notion when x is not in an "ambiguous area" (i.e., x is not very near a power of the radix, of larger than the largest representable number, or ±∞, or zero...);

—should be *useful*: after all, for a binary format with precision p, defining $\mathrm{ulp}\,(1)$ as 2^{-p} (i.e., $1-1^-$) or 2^{-p+1} (i.e., 1^+-1) are equally legitimate from a theoretical point of view. What matters is which choice is helpful (i.e., which choice will preserve in "ambiguous areas" properties that are true when we are far enough from them);

Let us consider the following common claims. They are true "in general", but they need some clarification. In the following $\mathrm{RN}\,(x)$ is x rounded to the nearest

[1] Partially only, because infinities and NaNs must also be defined.

[2] The possible implicit leading bit of the binary systems is counted in these p digits. For instance, in IEEE-754 double precision arithmetic, p is equal to 53.

(even) floating-point (FP) number, $\mathrm{RD}(x)$ is x rounded towards $-\infty$, $\mathrm{RU}(x)$ is x rounded towards $+\infty$, and $\mathrm{RZ}(x)$ is x rounded towards zero. The uppercase letter X will denote an FP number, whereas x will denote a real number.

Common claim 1.

$$X = RN(x) \Rightarrow |x - X| \leq \frac{1}{2}\, ulp$$

Common claim 2.

$$|x - X| < \frac{1}{2}\, ulp \Rightarrow X = RN(x)$$

or the following, slightly different variant,

$$|x - X| \leq \frac{1}{2}\, ulp \Rightarrow \textit{for all FP numbers } Y, |Y - x| \geq |X - x|.$$

Common claim 3.

$$|x - X| < 1\, ulp \Leftrightarrow X \in \{\, RD(x),\ RU(x)\}$$

In these claims, several things are unclear. The first one, of course, is the definition of ulp (especially near the powers of the radix). The second one is whether "ulp" means $\mathrm{ulp}(x)$ or $\mathrm{ulp}(X)$. Of course, in most practical cases, both values will be equal. But in difficult cases (e.g., X is a loose approximation to x, or these values are very close to a power of the radix), they may differ.

2. SHOULD WE CONSIDER ULP(EXACT) OR ULP(APPROXIMATION) ?

It should be clear that, for measuring the error of an approximation, the (possibly very loose) approximation should in general not be used for defining the measure of error: the distance between x (exact value) and X (FP approximation) should be expressed in terms of $\mathrm{ulp}(x)$, instead of $\mathrm{ulp}(X)$. Just consider the example given in Figure 1: we assume a binary floating-point system, with precision p (i.e., p-bit mantissas), we consider the real number $x = 1^+ = 1 + 2^{-p+1}$ and two (very poor) approximations $A = 2^- = 2 - 2^{-p+1}$ and $B = 2^+ = 2 + 2^{-p+2}$. A approximates x with error $(2^{p-1} - 2) \approx 2^{p-1}\,\mathrm{ulp}(A)$, whereas B approximates x with error $(2^{p-2} + 1/2) \approx 2^{p-2}\,\mathrm{ulp}(B)$. From these values, one could believe that B is a much better approximation to x than A. And yet, A is closer to x than B. This shows that ulp (approximation) cannot be a sensible unit for expressing errors.

3. VARIOUS DEFINITIONS OF FUNCTION ULP

Definition 2 KAHAN [754-R COMMITTEE 2004; KAHAN 2004]. *KahanUlp(x) is the width of the interval whose endpoints are the two finite representable numbers nearest x (even if x is not contained within that interval).*

Note: in [Harrison 1999], Harrison attributes the previous definition of $\mathrm{ulp}(x)$ to me, because I used approximately the same in my book on elementary functions [Muller 1997] (when writing the book, I was not aware of Kahan's definition).

Definition 3 HARRISON [HARRISON 1999]. *HarrisonUlp(x) is the distance between the closest straddling points a and b (i.e., those with $a \leq x \leq b$ and $a \neq b$), assuming an unbounded exponent range.*

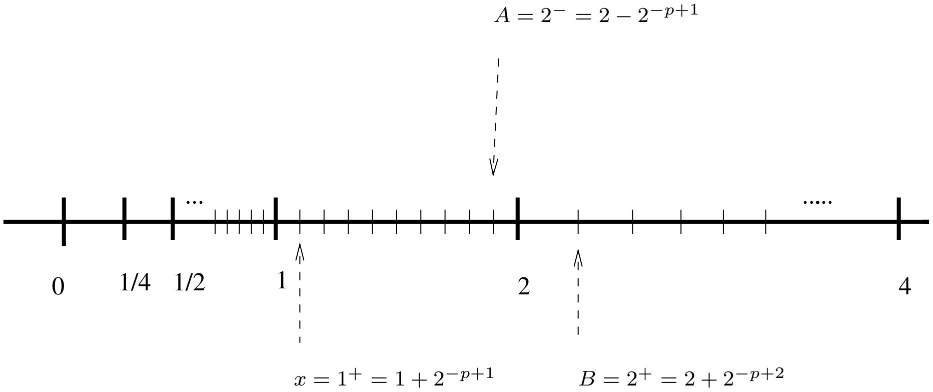
$A = 2^{-} = 2 - 2^{-p+1}$
...
.....
0
1/4
1/2
1
2
4
$x = 1^{+} = 1 + 2^{-p+1}$
$B = 2^{+} = 2 + 2^{-p+2}$

Fig. 1. *A approximates x with error $(2^{p-1} - 2) \approx 2^{p-1}$ ulp(A), whereas B approximates x with error $(2^{p-2} + 1/2) \approx 2^{p-2}$ ulp(B). From these values, one could believe that B is a much better approximation to x than A. And yet, A is closer to x than B.*

It is worth being noticed that Kahan's and Harrison's definitions coincide on FP numbers. However, for real numbers they may differ near powers of the radix. For instance, in radix 2 with precision p, if $1 < x < 1+2^{-p-1}$ then $\text{KahanUlp}(x) = 2^{-p}$ and $\text{HarrisonUlp}(x) = 2^{-p+1}$.

Definition 4 Goldberg [Goldberg 1991].
If the FP number $d_0.d_1d_2d_3d_4 \ldots d_{p-1}r^e$ is used to represent x, it is in error by

$$|d_0.d_1d_2d_3d_4 \ldots d_{p-1} - (x/r^e)|$$

units in the last place.

This definition uses the approximation that represents x: it does not define ulp as a function of x, since the value depends on which floating-point number approximates x. However, it clearly defines a function $\text{GoldbergUlp}(X)$, for a *floating-point* number $X \in [r^e, r^{e+1}]$, as r^{e-p+1}. Hence a natural generalization to real numbers is the following, that is equivalent to the one given by Cornea-Hasegan, Golliver and Markstein[3] [Cornea-Hasegan et al. 1999; Markstein 2000]: if $x \in [r^e, r^{e+1})$ then $\text{GoldbergUlp}(x) = r^{e-p+1}$.

One of the reviewers of this paper suggested the following definition.

Definition 5 Due to one of the reviewers.
Define a function $\mathcal{I}(x)$ as follows:

—$\mathcal{I}(0) = 0$;

—*between two consecutive floating point numbers X and Y, $X < Y$, $\mathcal{I}$ increases by 1, linearly.*

The ulp distance *between two real numbers a and b, denoted $\delta_{ulp}(a, b)$ is $|\mathcal{I}(b) - \mathcal{I}(a)|$.*

[3] They gave it in radix 2, but generalization to radix r is straightforward.

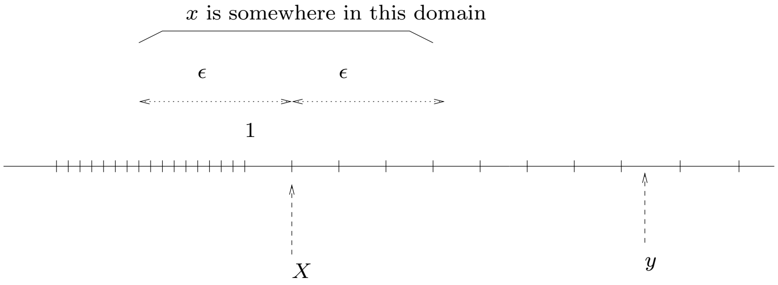

Fig. 2. With the ulp function suggested by one of the reviewers...

This definition is very elegant, and has nice properties for *expressing* errors (for instance, for saying that the maximum error of some elementary function is 0.501 ulp .), since with that definition, we always have

$$|x - X| < \frac{1}{2}\delta_{ulp}(x, X) \Rightarrow X = \mathrm{RN}\,(x)$$

and

$$|x - X| \leq \frac{1}{2}\delta_{ulp}(x, X) \Rightarrow \text{for all FP numbers } Y, |Y - x| \geq |X - x|.$$

and

$$|x - X| < 1\delta_{ulp}(x, X) \Leftrightarrow X \in \{\mathrm{RD}\,(x), \mathrm{RU}\,(x)\}\,.$$

Its drawback appears for actually *computing* error bounds: when one's calculations show that a computed result X is at some distance ϵ from the exact result x, one may want to express this error in terms of ulps.

Overton [Overton 2001] defines function ulp for FP numbers only. He defines ulp (X), for $X > 0$, as the gap between X and the next larger floating-point number (for $X < 0$, ulp $(X) =$ ulp $(-X)$). This value of ulp (X) is the same as $G(X)$, given above.

4. SOME PROPERTIES (ASSUMING UNBOUNDED EXPONENTS)

Definition 6. *A* regular *ulp function is such that there exists a value* $x_{cut} \in [1, 1 + r^{-p+1})$ *so that*

$$ulp\,(x) = r^{-p+1+k}$$

if $r^k x_{cut} < x < r^{k+1} x_{cut}$. *The number* x_{cut} *will be called the "cutting point" of the ulp function.*

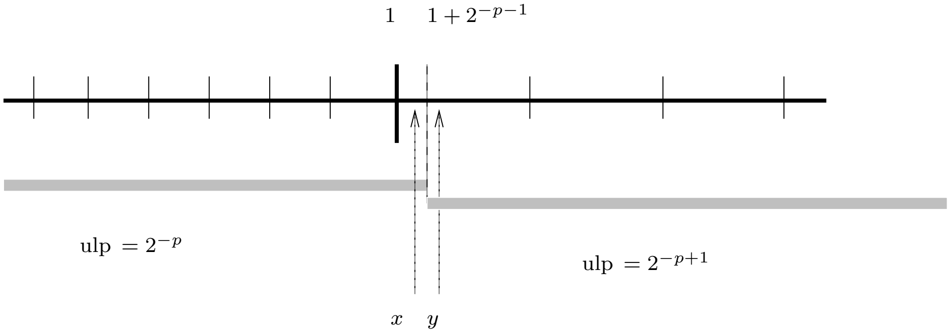

Fig. 3. *The values of KahanUlp(x) near 1, assuming a binary FP system with precision p. Note the strange side effect: 1 seems to be a better approximation to y than to x.*

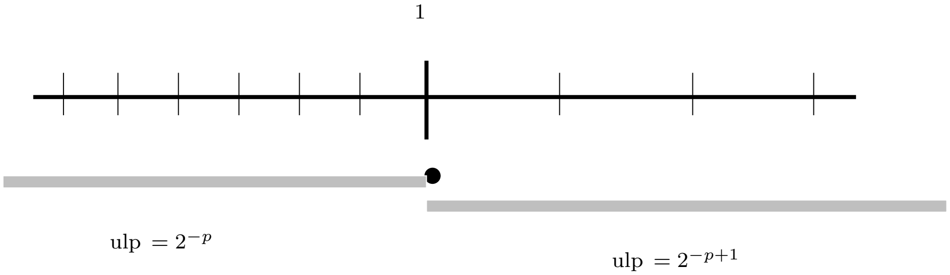

Fig. 4. *The values of HarrisonUlp(x) near 1, assuming a binary FP system with precision p.*

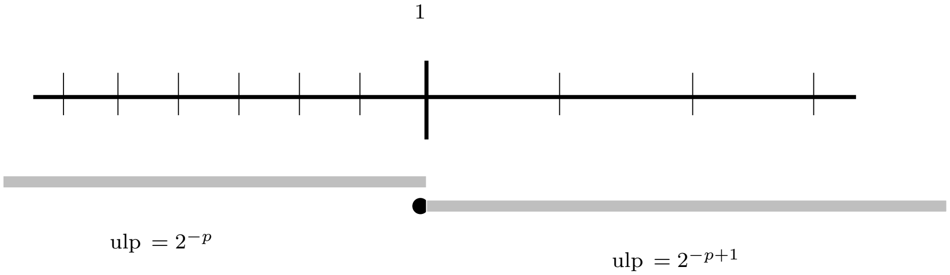

Fig. 5. *The values of Modified GoldbergUlp(x) near 1, assuming a binary FP system with precision p. Notice that Modified GoldbergUlp(x) and HarrisonUlp(x) only differ when x is a power of the radix.*

This does not uniquely define the value of $\mathrm{ulp}\,(x)$ since there remains an ambiguity at $x = r^k x_{\mathrm{cut}}$. This ambiguity has no importance if $x_{\mathrm{cut}} \neq 1$, but may make a difference if $x_{\mathrm{cut}} = 1$.

For instance, both HarrisonUlp and KahanUlp are regular ulp functions, with $x_{\mathrm{cut}} = 1$ for HarrisonUlp and $x_{\mathrm{cut}} = 1 + \frac{r^{-p}}{2}(r-1)$ for KahanUlp.

4.1 With rounding to nearest

Theorem 1. *We have*

$$|X - x| < \frac{1}{2}\, ulp\,(x) \Rightarrow X = RN(x)$$

for any real x and FP number X, if and only if

$$x_{cut} \geq 1 + r^{-p}(\frac{r}{2} - 1). \tag{1}$$

Proof: We only consider the case $1 \leq x < 1^+$ (the other cases are either straightforward, or easily deduced from this one). First, if $x > x_{\mathrm{cut}}$, then $\mathrm{ulp}\,(x) = r^{-p+1}$. In that case, since $1^- = 1 - r^{-p}$ cannot be the FP number that is nearest x (because x is closer to 1 than to 1^-), we must have

$$x - 1^- \geq \frac{1}{2}\,\mathrm{ulp}\,(x),$$

i.e.,

$$x \geq 1 + r^{-p}(\frac{r}{2} - 1).$$

This gives the condition of the theorem.

Conversely, if $x_{\mathrm{cut}} \geq 1 + r^{-p}(\frac{r}{2} - 1)$ then

—if $1 \leq x < x_{\mathrm{cut}}$ then $\mathrm{ulp}\,(x) = 1 - 1^- = r^{-p}$. Hence, the only values that can be within $\frac{1}{2}\,\mathrm{ulp}\,(x)$ from x (if any) are 1 and 1^+, and at most one of these values only can be within $\frac{1}{2}\,\mathrm{ulp}\,(x)$ from x. If there is one, it will necessary be the FP number that is nearest x;

—if $x > x_{\mathrm{cut}}$ then $\mathrm{ulp}\,(x) = 1^+ - 1 = r^{-p+1}$. Since (1) implies that $x - 1^- > \frac{1}{2}\,\mathrm{ulp}\,(x)$, the only values that can be within $\frac{1}{2}\,\mathrm{ulp}\,(x)$ from x (if any) are 1 and 1^+, and at most one of these values only can be within less than $\frac{1}{2}\,\mathrm{ulp}\,(x)$ from x. If there is one, it will necessary be the FP number that is nearest x.

Theorem 2. *To have*

$$X = RN(x) \Rightarrow |X - x| \leq \frac{1}{2}\, ulp\,(x)$$

for any real x and FP number X, we need

$$x_{cut} \leq 1 + \frac{1}{2} r^{-p}. \tag{2}$$

Proof: Again, we only consider the case $1 \leq x < 1^+$ (the other cases are either straightforward, or easily deduced from this one).

If $x_{\mathrm{cut}} > 1 + \frac{1}{2} r^{-p}$ then, for

$$1 + \frac{1}{2} r^{-p} < x < \min\{x_{\mathrm{cut}}, 1 + \frac{1}{2} r^{-p+1}\}$$

we have,

$$\begin{cases} \mathrm{RN}\,(x) &= 1 \\ \mathrm{ulp}\,(x) &= r^{-p} \end{cases}$$

hence, we have $1 = \mathrm{RN}(x)$, and yet $|1-x| > \frac{1}{2}\,\mathrm{ulp}(x)$. Hence the condition of the theorem.

Conversely, if $x_{\mathrm{cut}} \leq 1 + \frac{1}{2}r^{-p}$, then for $1 \leq x \leq x_{\mathrm{cut}}$, we have both $\mathrm{RN}(x) = 1$ and $|1-x| \leq \frac{1}{2}\,\mathrm{ulp}(x)$, and for $x_{\mathrm{cut}} < x < 1^+$, we have $\mathrm{ulp}(x) = r^{-p+1} = 1^+ - 1$, so $\mathrm{RN}(x)$ is the value X in $\{1, 1^+\}$ that is nearest x, and $|X - x|$ is obviously less than or equal to $(1^+ - 1)/2 = \frac{1}{2}\,\mathrm{ulp}(x)$.

Property 1. *In radix* 2*,*

$$|X - x| < \frac{1}{2}\,\mathit{HarrisonUlp}(x) \Rightarrow X = \mathit{RN}(x)$$

See Theorem 1 for proof. Property 1 is not true in radices greater than or equal to 3. Figure 6 gives a counter-example in radix 3.

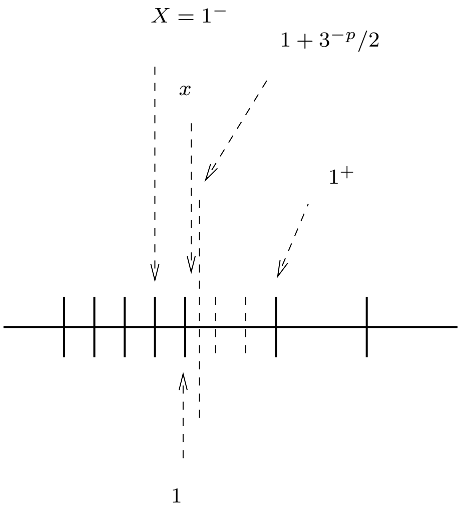

Fig. 6. *This example shows that Property 1 is not true in radix* 3*. Here, x satisfies $1 < x < 1 + \frac{1}{2}3^{-p}$ and $X = 1^- = 1 - 3^{-p}$. We have $\mathit{HarrisonUlp}(x) = 3^{-p+1}$, and $|x - X| < 3^{-p+1}/2$, so that $|x - X| < \frac{1}{2}\,\mathit{HarrisonUlp}(x)$. And yet, $X \neq \mathit{RN}(x)$.*

Property 2. *For any radix,*

$$X = \mathit{RN}(x) \Rightarrow |X - x| \leq \frac{1}{2}\,\mathit{HarrisonUlp}(x)$$

See Theorem 2 for proof.

Property 3. *For any radix,*

$$|X - x| < \frac{1}{2}\,\mathit{KahanUlp}(x) \Rightarrow X = \mathit{RN}(x)$$

See Theorem 1 for proof.

Property 4. *In radix* 2,

$$X = RN(x) \Rightarrow |X - x| \leq \frac{1}{2}\, KahanUlp\,(x)$$

See Theorem 2 for proof. Property 4 is not true in radices greater than or equal to 3. Figure 7 gives a counter-example in radix 3.

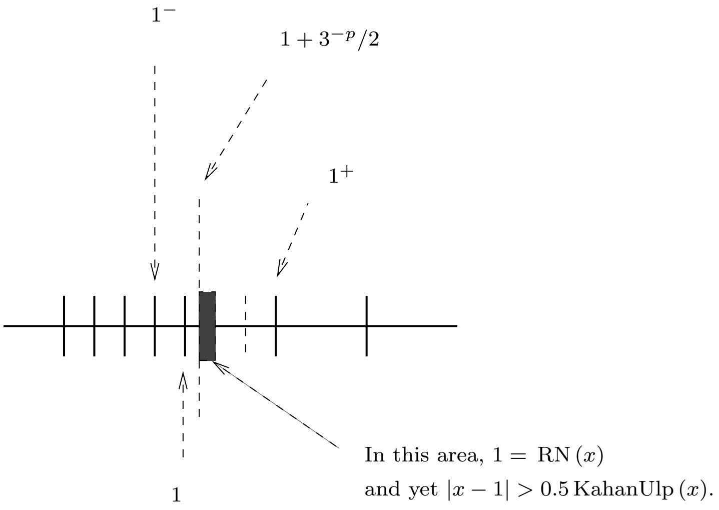
1^-
$1 + 3^{-p}/2$
1^+
1
In this area, $1 = \mathrm{RN}\,(x)$
and yet $|x - 1| > 0.5\,\mathrm{KahanUlp}\,(x)$.

Fig. 7. *This example shows that Property 4 is not true in radix* 3. *If* $1 + \frac{1}{2}3^{-p} < x < 1 + 3^{-p}$, *then* $1 = RN(x)$, *and yet* $|x - 1| > 0.5\, KahanUlp\,(x)$.

We see that with rounding to nearest in radix 2, both Kahan's and Harrison's definitions preserve the common claims listed above. As we shall see later, the situation is different with directed roundings.

Property 5. *In radix* 2,

$$|X - x| < \frac{1}{2}\, GoldbergUlp\,(x) \Rightarrow X = RN(x)$$

Property 6. *For any radix,*

$$X = RN(x) \Rightarrow |X - x| \leq \frac{1}{2}\, GoldbergUlp\,(x)$$

See Theorem 2 for proof.

Theorem 3. *If the radix r is greater than or equal to* 4, *there is no regular ulp function that satisfies both*

$$|X - x| < \frac{1}{2}\, ulp\,(x) \Rightarrow X = RN(x)$$

and

$$X = RN(x) \Rightarrow |X - x| \leq \frac{1}{2}\, ulp\,(x).$$

Theorem 3 implies that for $r \geq 4$ (which means, in practice, for $r = 10$, since radices different from 2 and 10 seem no longer used) we have to choose between both properties: they will be true "in general", but at least one of them will be wrong when x is very close to a power of r. Theorem 3 is an immediate consequence of Theorems 1 and 2 (conditions (1) and (2) become incompatible for $r \geq 4$). For $r = 3$, the only allowable value of x_{cut} is $1+3^{-p}/2$. For $r = 2$, $x_{\text{cut}} \in [1, (1+1^+)/2]$.

4.2 With directed roundings

Property 7. *For any value of the radix r,*

$$X \in \{\, RD\,(x),\ RU\,(x)\} \Rightarrow |X - x| < 1\, HarrisonUlp\,(x)$$

But now the converse is not true. There are values X and x for which $|X - x| < 1\,\text{HarrisonUlp}\,(x)$, and yet X is not in $\{\,\text{RD}\,(x),\ \text{RU}\,(x)\}$ (consider the case x slightly above 1 and X equal to 1^-, the FP predecessor of 1).

With $\text{KahanUlp}\,(x)$, also, there are values X and x for which $|X-x| < 1\,\text{HarrisonUlp}\,(x)$, and yet X is not in $\{\,\text{RD}\,(x),\ \text{RU}\,(x)\}$ (consider, in radix 2 with precision p, the case $X = 1 - 2^{-p}$ and x between $1 + 2^{-p-1}$ and $1 + 2^{-p}$).

With $\text{KahanUlp}\,(x)$, there is no equivalence of property 7. As noticed by Harrison [Harrison 1999], we can have $X \in \{\,\text{RD}\,(x),\ \text{RU}\,(x)\}$, and $|X - x|$ significantly larger than $1\,\text{KahanUlp}\,(x)$ (it can be arbitrarily close, without being equal, to $r\,\text{KahanUlp}\,(x)$). Consider the radix-2 case depicted by Figure 8.

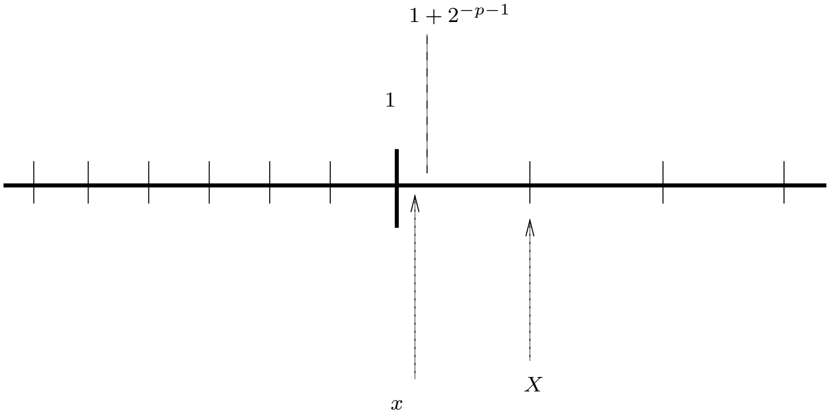

Fig. 8. *We assume radix* 2 *and precision p. X is equal to* $RU(x)$*, and yet* $|X - x|$ *is very close to* $2\,KahanUlp\,(x)$ *[Harrison 1999].*

4.3 If anyway one decides to use ulp (X)

Although we have indicated in Section 2 that using $\text{ulp}\,(x)$ as the measure of error seems much preferable, one may, for some application, find a good reason for using $\text{ulp}\,(X)$. In such a case, we list the obtained properties below.

Property 8. *Assuming unbounded exponents, we find, for any value of the radix r:*

$$
\begin{gathered}
|X-x| < \tfrac{1}{2}\,\mathit{HarrisonUlp}\,(X) \Rightarrow X = \mathit{RN}(x)\\
X = \mathit{RN}(x) \text{ does not imply } |X-x| \le \tfrac{1}{2}\,\mathit{HarrisonUlp}\,(X)\\
|X-x| < \mathit{HarrisonUlp}\,(X) \Rightarrow X \in \{\,\mathit{RD}(x), \mathit{RU}(x)\}\\
X \in \{\,\mathit{RD}(x), \mathit{RU}(x)\} \text{ does not imply } |X-x| \le \mathit{HarrisonUlp}\,(X)\\
|X-x| < \tfrac{1}{2}\,\mathit{KahanUlp}\,(X) \Rightarrow X = \mathit{RN}(x)\\
X = \mathit{RN}(x) \text{ does not imply } |X-x| \le \tfrac{1}{2}\,\mathit{KahanUlp}\,(X)\\
|X-x| < \mathit{KahanUlp}\,(X) \Rightarrow X \in \{\,\mathit{RD}(x), \mathit{RU}(x)\}\\
X \in \{\,\mathit{RD}(x), \mathit{RU}(x)\} \text{ does not imply } |X-x| \le \mathit{KahanUlp}\,(X)\\
X = \mathit{RN}(x) \Rightarrow |X-x| \le \tfrac{1}{2}\,\mathit{GoldbergUlp}\,(X)\\
|X-x| < \tfrac{1}{2}\,\mathit{GoldbergUlp}\,(X) \text{ does not imply } X = \mathit{RN}(x)\\
X \in \{\,\mathit{RD}(x), \mathit{RU}(x)\} \Rightarrow |X-x| \le \mathit{GoldbergUlp}\,(X)\\
|X-x| < \mathit{GoldbergUlp}\,(X) \text{ does not imply } X \in \{\,\mathit{RD}(x), \mathit{RU}(x)\}
\end{gathered}
$$

In that case, Kahan's and Harrison's definitions satisfy the same properties, which is not surprising since they coincide on FP numbers.

5. LINK BETWEEN RELATIVE ERROR AND ERROR EXPRESSED IN ULPS

If X is an approximation to some real number x, its accuracy can be measured by the relative error (see for instance [Higham 2002]).

$$\left|\frac{x-X}{x}\right|.$$

Relative errors are very often used to express the accuracy of numerical computations. Let us examine how to convert an error expressed in ulps from and to a relative error.

Assume an ulp function, with cutting-point x_{cut}. Assume the floating-point number X approximates a nonzero real number x. Assume $2^k \le |x| < 2^{k+1}$.

5.1 Conversion from relative errors to ulps

Assume

$$\left|\frac{x-X}{x}\right| = \epsilon_r. \tag{3}$$

If the mantissa of x is less than x_{cut} then $\mathrm{ulp}\,(x) = r^{k-p}$, hence (3) implies

$$|x-X| = \epsilon_r |x| r^{p-k}\,\mathrm{ulp}\,(x),$$

which implies

$$|x-X| \le \epsilon_r x_{\text{cut}} r^p\,\mathrm{ulp}\,(x).$$

If the mantissa of x is larger than x_{cut}, a similar calculation gives

$$|x-X| \le \epsilon_r r^p\,\mathrm{ulp}\,(x).$$

Therefore, a relative error ϵ_r implies an error in ulps bounded by

$$|x - X| \leq \epsilon_r x_{\text{cut}} r^p \,\mathrm{ulp}\,(x). \tag{4}$$

This can be (very) slightly larger than the usually assumed bound $\epsilon_r \times r^p$. For instance, with Kahan's definition in radix-2, we get the bound

$$|x - X| \leq \epsilon_r \left(2^p + \frac{1}{2}\right) \mathrm{ulp}\,(x).$$

In practice, we can get very close to that bound. For instance, if $x = 1 + 2^{-p-1} - 2^{-2p}$ (that is, x is very slightly below x_{cut}) and $X = 1$. The relative error is

$$2^{-p-1} - 5 \times 2^{-2p-2} + 9 \times 2^{-3p-3} - \cdots$$

and the error in ulps is

$$\left(\frac{1}{2} - 2^{-p}\right) \mathrm{KahanUlp}\,,$$

hence the error in ulps is approximately

$$2^p + \frac{1}{2} - 2^{-p}$$

times the relative error.

5.2 Conversion from ulps to relative errors

Assume that $|x - X| = \alpha \,\mathrm{ulp}\,(x)$. We easily get:

$$\left|\frac{x - X}{x}\right| \leq \alpha \times r^{-p+1}, \tag{5}$$

which is the bound that is usually assumed.

6. PROPERTIES NEAR INFINITY

Kahan's definition clearly defines function ulp for big numbers. Harrison's definition too, since it assumes unbounded exponents, but in a way that does not allow to preserve claims 1, 2 and 3. Define L as the largest finite FP number, and L^- as its predecessor. If x is larger than L, then it is clear from definition 2 that

$$\mathrm{KahanUlp}\,(x) = L - L^-.$$

From this, for big x, it is clear that

$$|X - x| < \frac{1}{2} \mathrm{KahanUlp}\,(x) \Rightarrow X = \mathrm{RN}\,(x)$$

So, property 3 is always true (there is no need to assume unbounded exponents, as in the previous section).

Interestingly enough, with IEEE-754 FP (binary) numbers, the converse holds. This is due to a feature of the IEEE-754 Standard [American National Standards Institute and Institute of Electrical and Electronic Engineers 1985] (which by the way makes $\mathrm{RN}\,(x)$ quite different from what one would expect from the term "rounding to *nearest*"). The standard says that an infinitely precise result with magnitude at least

$$2^{\mathrm{emax}}\left(2-2^{-p}\right)$$

shall round to ∞ with no change in sign. With that convention, if X is *finite*,

$$X = \mathrm{RN}\,(x) \Rightarrow |X-x| \leq \frac{1}{2}\,\mathrm{KahanUlp}\,(x),$$

i.e., Property 4 remains true for big numbers.

7. CONCLUSION

It appears that a definition that would preserve most properties would be

$$\mathrm{ulp}\,(x) = \begin{cases} \mathrm{HarrisonUlp}\,(x) & \text{if } |x| \leq L \\ \mathrm{KahanUlp}\,(x) = L - L^{-} & \text{otherwise,} \end{cases}$$

which could be given as follows:

Definition 7. *If x is a real number that lies between two finite consecutive FP numbers a and b, without being equal to one of them, then ulp $(x) = |b-a|$, otherwise ulp (x) is the distance between the two finite FP numbers nearest x. Moreover, ulp (NaN) is NaN.*

Acknowledgement

In November 2004, I had the opportunity to discuss these topics (as well as other aspects of floating-point arithmetic) with Professor Kahan. These were enlightening discussions. Although we may occasionally disagree on minor issues, we share the same feeling that the big improvements brought to numerical computing by the IEEE-754 and 854 standards for floating-point arithmetic are endangered: we must explain to computer architects, compiler designers and numerical application programmers that some features of the standards that sometimes seem arcane or that seem to hinder performance may be crucial when reliability and/or portability are at stake.

REFERENCES

754-R Committee. 2004. DRAFT standard for floating-point arithmetic p754 d0.6.5. Available at `http://www.validlab.com/754R/drafts/754r.pdf`.

American National Standards Institute and Institute of Electrical and Electronic Engineers. 1985. IEEE standard for binary floating-point arithmetic. *ANSI/IEEE Standard, Std 754-1985,* New York.

Cornea-Hasegan, M. A., Golliver, R. A., and Markstein, P. 1999. Correctness proofs outline for Newton-Raphson based floating-point divide and square root algorithms. In *Proceedings of the 14th IEEE Symposium on Computer Arithmetic (Adelaide, Australia)*, Koren and Kornerup, Eds. IEEE Computer Society Press, Los Alamitos, CA, 96–105.

Goldberg, D. 1991. What every computer scientist should know about floating-point arithmetic. *ACM Computing Surveys 23,* 1 (Mar.), 5–47.

Harrison, J. 1999. A machine-checked theory of floating-point arithmetic. In *Theorem Proving in Higher Order Logics: 12th International Conference, TPHOLs'99*, Y. Bertot, G. Dowek, A. Hirschowitz, C. Paulin, and L. Théry, Eds. Lecture Notes in Computer Science, vol. 1690. Springer-Verlag, Berlin, 113–130.

Higham, N. 2002. *Accuracy and Stability of Numerical Algorithms, Second Edition.* SIAM, Philadelphia, PA.

Kahan, W. 1996. Lecture notes on the status of IEEE-754. PDF file accessible electronically through the Internet at the address `http://www.cs.berkeley.edu/~wkahan/ieee754status/IEEE754.PDF`.

Kahan, W. 2004. A logarithm too clever by half. Available at `http://http.cs.berkeley.edu/~wkahan/LOG10HAF.TXT`.

Markstein, P. 2000. *IA-64 and Elementary Functions : Speed and Precision.* Hewlett-Packard Professional Books. Prentice Hall, Englewood Cliffs, NJ.

Muller, J. 1997. *Elementary Functions, Algorithms and Implementation.* Birkhauser, Boston.

Overton, M. A. 2001. *Numerical Computing with IEEE Floating-Point Arithmetic.* SIAM, Philadelphia, PA.

Appendix: Maple programs that compute ulp (x) in double precision

The following two Maple programs compute $\mathrm{KahanUlp}\,(t)$ and $\mathrm{ulp}\,(t)$ as suggested in Definition 7 for any real number t, assuming that the used floating-point format is the double precision format of the IEEE-754 standard (i.e., $r = 2$ and $p = 53$S).

```
KahanUlp := proc(t);
x := abs(t);
if x < 2^(-1021) then res := 2^(-1074)
   else if x > (1-2^(-53))*2^(1024) then res := 2^971
   else
     powermin := 2^(-1021); expmin := -1021;
     powermax := 2^1024; expmax := 1024;
# x is between powermin = 2^expmin and powermax = 2^expmax
     while (expmax-expmin > 1) do
       expmiddle := round((expmax+expmin)/2);
       powermiddle := 2^expmiddle;
       if x >= powermiddle then
             powermin := powermiddle;
             expmin := expmiddle
       else
             powermax := powermiddle;
             expmax := expmiddle
       fi;
      od;
# now, expmax - expmin = 1
# and powermin <= x < powermax
       if x/powermin <= 1+2^(-54) then res := 2^(expmin-53)
        else res := 2^(expmin-52)
       fi;
    fi;
fi;
res;
end;
```

```
SuggestedUlp := proc(t);
x := abs(t);
if x < 2^(-1021) then res := 2^(-1074)
   else if x > (1-2^(-53))*2^(1024) then res := 2^971
   else
```

```
     powermin := 2^(-1021); expmin := -1021;
     powermax := 2^1024; expmax := 1024;
# x is between powermin = 2^expmin and powermax = 2^expmax
     while (expmax-expmin > 1) do
       expmiddle := round((expmax+expmin)/2);
       powermiddle := 2^expmiddle;
       if x >= powermiddle then
             powermin := powermiddle;
             expmin := expmiddle
       else
             powermax := powermiddle;
             expmax := expmiddle
       fi;
      od;
# now, expmax - expmin = 1
# and powermin <= x < powermax
       if x = powermin then res := 2^(expmin-53)
        else res := 2^(expmin-52)
       fi;
    fi;
fi;
res;
end;
```