ULPS AND RELATIVE ERROR

Marius Cornea

24th IEEE Symposium on Computer Arithmetic

July 25th, 2017

What is an ULP? (Unit-in-the-Last-Place)

William Kahan, *A logarithm too clever by half*, 1960: ulp(x) is the gap between the two floating-point (FP) numbers nearest to x, even if x is one of them

The IEEE Standard 754-1985 mentioned 'units in the last place', but did not define them

The IEEE Standard 754-2008 defines the quantum of a FP number:

- The quantum of a finite floating-point representation is the value of a unit in the last position of its significand. This is equal to the radix raised to the exponent q; the significand is regarded as an integer

J.M. Muller & al., Handbook of Floating-Point Arithmetic defines ULP (& uses it 256x on 625 pages)

- Defines ulp as a function, 'closely related to the notion of quantum'
- J. Harrison, *A machine-checked theory of floating point arithmetic*: ulp(x) is the distance between the closest straddling floating-point numbers a and b (i.e., with $a \le x \le b$ and $a \ne b$), assuming the exponent not upper-bounded
- D. Goldberg, *What Every Computer Scientist…*: If the FP number $d_0.d_1d_2d...d_{p-1}\ \beta^e$ is used to represent x, it is in error by $|\ d_0.d_1d_2d...d_{p-1} - x/\beta^e|$ units in the last place. For $\beta=2$ and $f = \sigma \cdot s_N \cdot 2^e$, 1 ulp = 2^{e-N+1}
- Cornea, Golliver, and Markstein, *Correctness proofs outline for Newton–Raphson based floating-point divide and square root algorithms*: (Goldberg's definition, extended to reals): If $|x| \in [\beta^e,\beta^{e+1})$, then $\text{ulp}(x) = \beta^{\max(e,emin)-p+1}$

Relative Error vs. ULPs, and Conversion Rules

Relative error: absolute error divided by the exact value, $(f–x)/x$ (sign is preserved!)

Ulp error: absolute error div. by the weight of the LSB of the FP approx., $(f–x)/2^{e–N+1}$

The MPFR Team, *The MPFR Library: Algorithms and Proofs* (n: the working precision)

- If $x = m \cdot 2^e$ with $1/2 \le |m| < 1$ and $e := \exp(x)$, then $\mathrm{ulp}(x)$ is defined as $\mathrm{ulp}(x) := 2^{\exp(x)-n}$
- If $u = \circ(x)$, then the error on u is at most $\mathrm{ulp}(u) = 2^{\exp(u)-n} \le |u| \cdot 2^{1-n}$; the relative error is $\le 2^{1-n}$
- If the relative error is $\le \delta$, then the [ulp] error is at most $\delta \cdot |u| \le \delta \cdot 2^n \mathrm{ulp}(u)$
- Going from the ulp-error to the relative error and back, we lose a factor of two

Jean-Michel Muller & al, *Handbook of Floating-Point Arithmetic*:

- "It is important … to establish links between errors expressed in ulps, and relative errors"
- Converting between errors in ulps and relative errors (x is a real, X is a floating-point value):
  - Assume that $|x - X| = \alpha \cdot \mathrm{ulp}(x)$. Assuming no underflow, $|(x – X)/x| \le \alpha \cdot \beta^{-p+1}$
  - A relative error $\varepsilon_r = |(x-X)/x|$ implies an error in ulps $|x-X| \le \varepsilon_r \cdot \beta^p \cdot \mathrm{ulp}(x)$
- These relationships allow for easier proofs in numerical analysis, e.g. in cases where we need to prove that certain functions or FP operators yield correctly rounded results

David Goldberg, *What Every Computer … :* similar conversion rules, exemplified for 1/2 ulp

Then What Are the Contributions of This Paper?

- Explained (hopefully better!) how to work with ulps in all situations, e.g. when values of concern cross the boundary between two consecutive binades
- Extended the definition of the ulp to two real values (or rather, specifies which one to use)
- Established equivalence relations between errors expressed in ulps and relative errors (not at binade level)
- Maintained the sign in the relative and ulp errors
- Determined one property (Theorem 1) which was used to derive several others (including those shown from current literature)
- Determined (slightly) tighter error bounds in some cases
- Put it all in one place…

A Few Definitions

The following definitions have meaning only in the context of operating with floating-point numbers having N-bit significands, with N specified (the 'precision')

D1. Approximation of a real number: floating-point number $f = \sigma \cdot s \cdot 2^e$ that approximates the real number x with a specified relative error, or is within a specified number of ulps of x

D2. Ulp, or unit in the last place associated with a floating-point number f: the weight of the least significant bit in the significand of a floating-point number; the exponent is e, then one ulp has magnitude 2^{e-N+1}; also written as ulp(f)

D3. Ulp associated with a real number x: if $2^e \le x < 2^{e+1}$, then one ulp has magnitude 2^{e-N+1}; also written as ulp(x)

D4. Ulp associated with a pair of real numbers x and y: it is the smaller of ulp(x) and ulp(y), i.e. min(ulp(x), ulp(y)); also written as ulp(x,y)

A Few Definitions (continued)

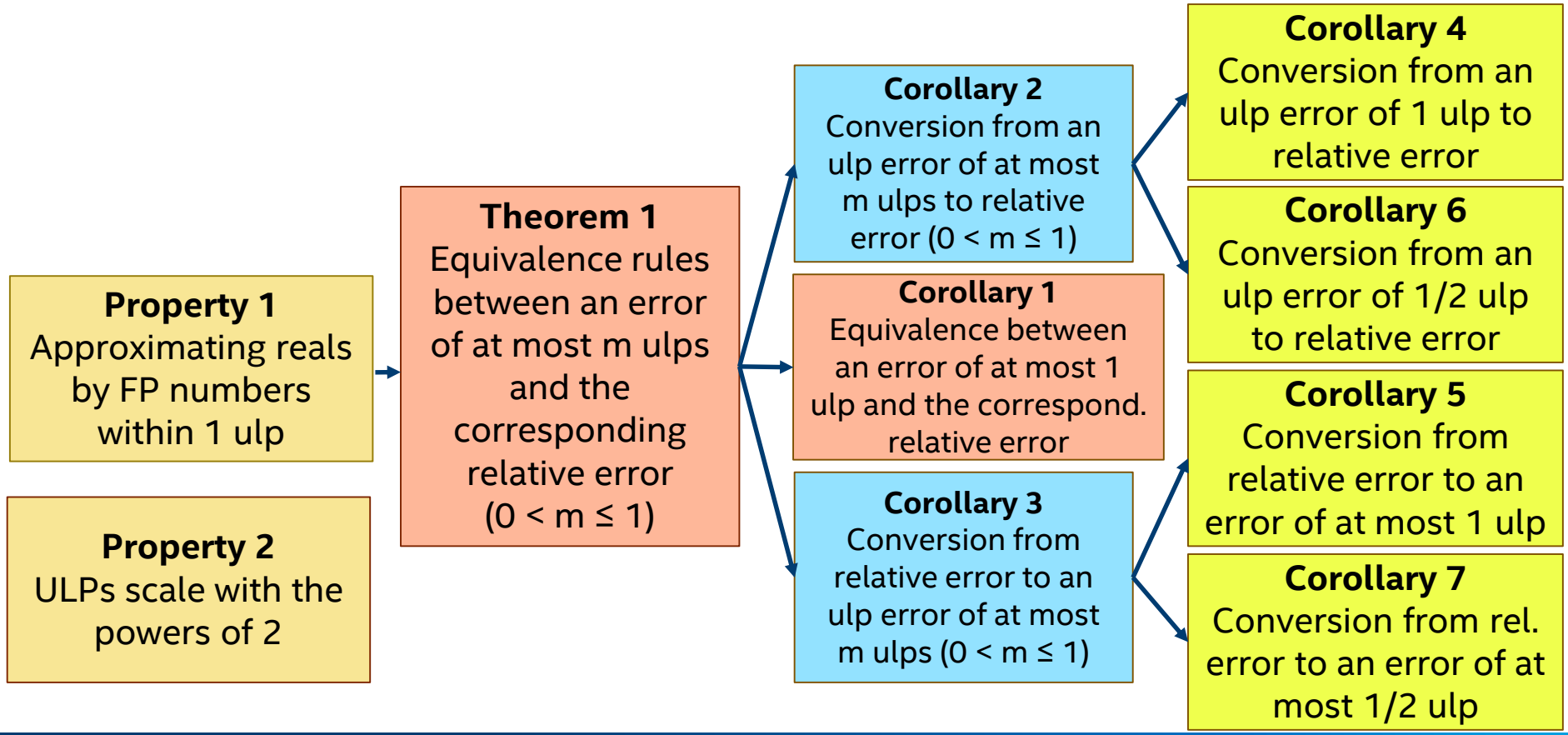
Note that from here on, whenever we say “f ≈ x within m ulps of x” or “f approximates x within m ulps of x” where f, x ϵ **R**, by ‘ulp’ we mean ulp(f,x)

D5. y is within 1 ulp of x: the real number y is "close" to the real number x such that $|x – y| < 1$ ulp(x,y); if the relation is non-strict, this becomes $|x – y| \leq 1$ ulp(x,y) (we consider [mostly] a strict relationship in the paper)

D6. f approximates x, within one ulp of x: also written as f ≈ x, within 1 ulp of x; the floating-point number f is "close" to the real number x, such that $|x – f| < 1$ ulp;

- $f = 1.0 \cdot 2^e$ can approximate real numbers both to its right and to its left; it is assumed that $f \cdot x > 0$; note also that if the relation is non-strict, the inequality above becomes $|x – f| \leq 1$ ulp

Relationship of Properties from this Paper



Theorem 1 (and only…)

Notations: Unless specified otherwise, $x \in \mathbf{R}^*$ is a real non-zero number, and $f \in \mathbf{Q}^*$ is a floating-point non-zero number with N bits in the significand

$$f = \sigma \cdot s \cdot 2^e, \quad \sigma = \pm 1, s = 1 + k/2^{N-1}, k \in \mathbf{Z}, 0 \le k \le 2^{N-1} - 1, e \in \mathbf{Z} \text{ (unbounded)}$$

Theorem 1 (ULPs versus Relative Error)

Let f and x as defined above, and $m \in \mathbf{R}$, $0 < m \le 1$. Let ε be the relative error when approximating x by f, $\varepsilon = (f - x)/x$.

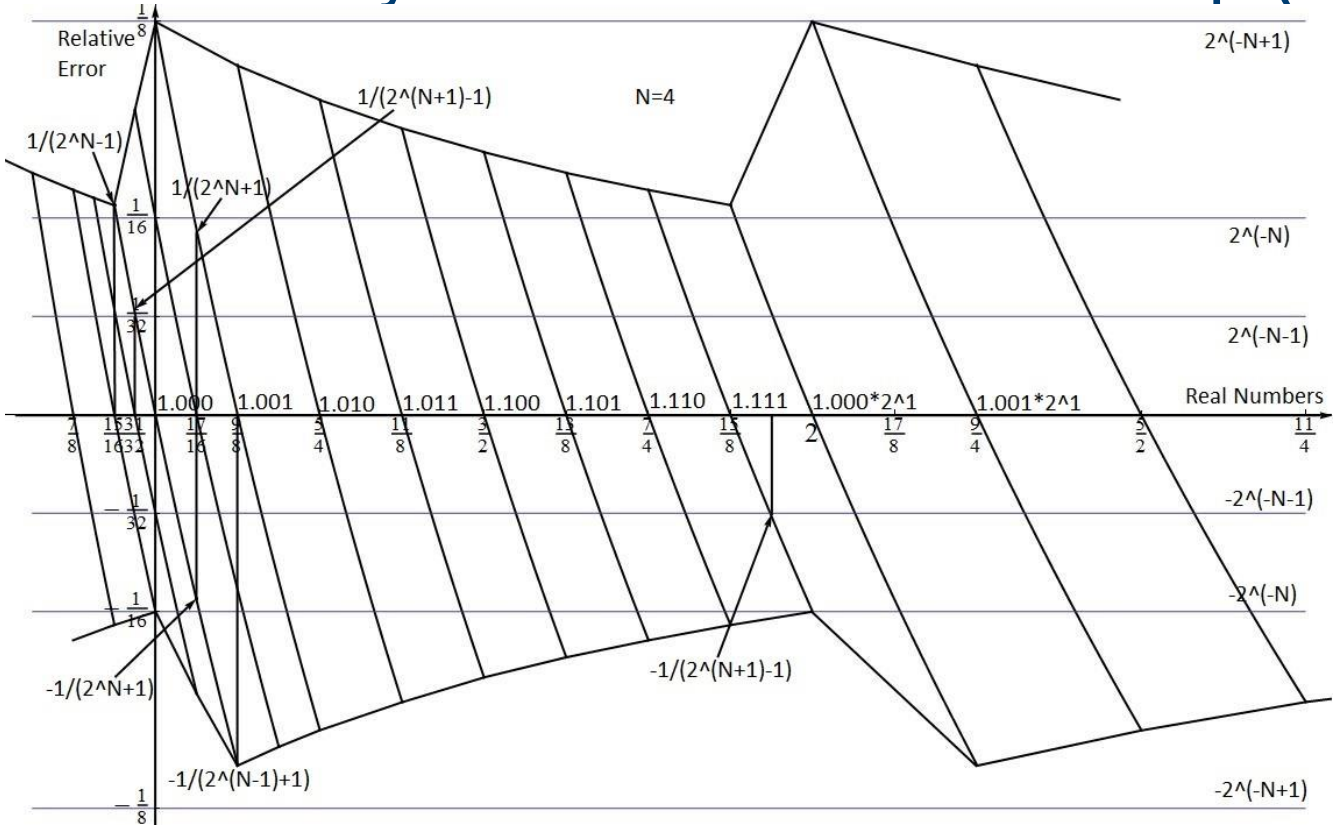
(a) For $k \neq 0$, when $f = \sigma \cdot (1 + k/2^{N-1}) \cdot 2^e$, $1 \le k \le 2^{N-1} - 1$:

$f \approx x$ within m ulps of $x \Leftrightarrow f = x \cdot (1 + \varepsilon), \quad \varepsilon \in (-m/(2^{N-1} + k + m), m/(2^{N-1} + k - m))$

(b) For $k = 0$, when $f = \sigma \cdot 2^e$:

$f \approx x$ within m ulps of $x \Leftrightarrow f = x \cdot (1 + \varepsilon), \varepsilon \in (-m/(2^{N-1} + m), m/(2^N - m))$

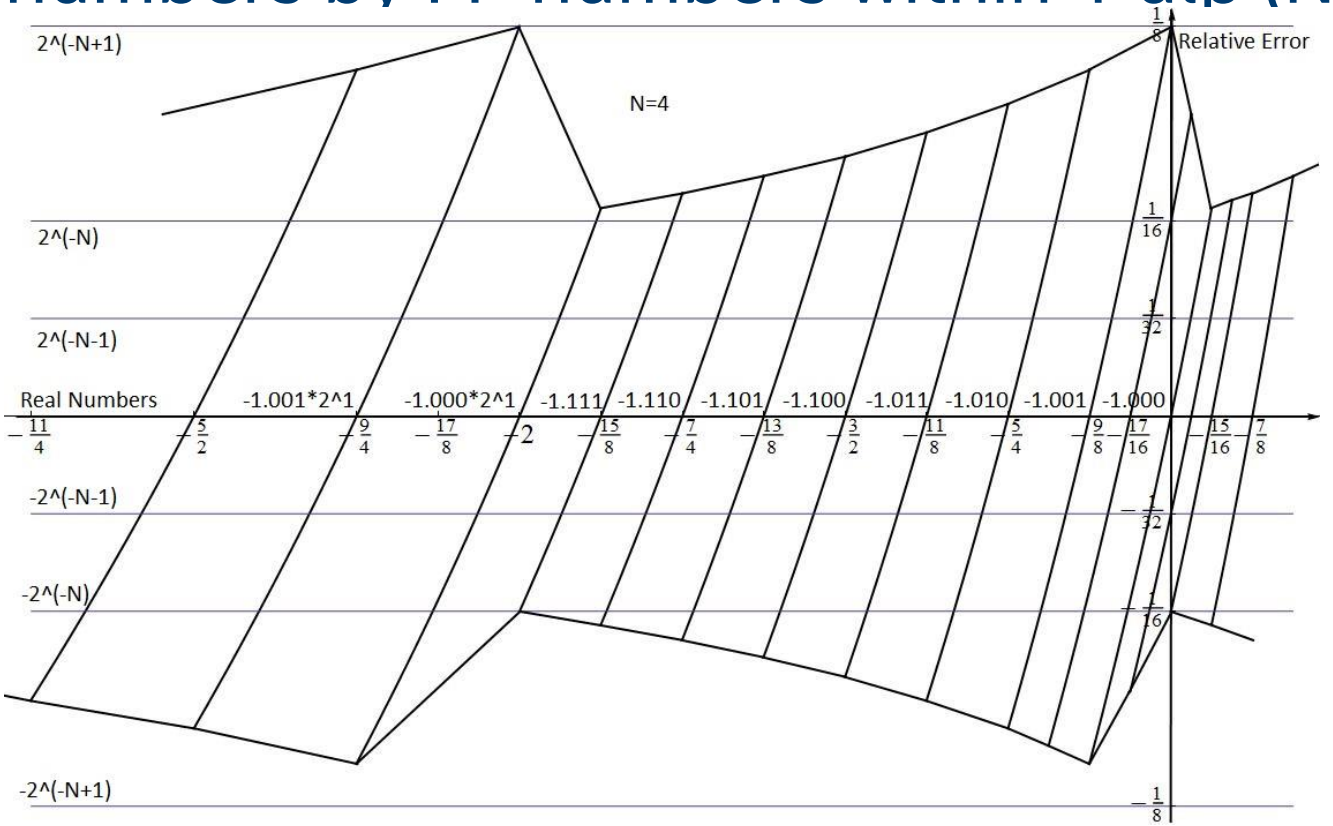
Relative error when approximating positive real numbers by FP numbers within 1 ulp (N=4)



Outer envelope:

maximum relative error when approximating positive real numbers by other real numbers within 1 ulp.

Relative error when approximating negative real numbers by FP numbers within 1 ulp (N=4)



Outer envelope:

maximum relative error when approximating negative real numbers by other real numbers within 1 ulp.

Corollaries

Corollary 1

Let f and x as defined above. Then:

(a) For $k \neq 0$, when $f = \sigma \cdot (1 + k/2^{N-1}) \cdot 2^e$, $1 \leq k \leq 2^{N-1} - 1$:

$f \approx x$ within 1 ulp of $x \Leftrightarrow f = x \cdot (1 + \varepsilon)$, $\varepsilon \in (-1 / (2^{N-1} + k + 1), 1 / (2^{N-1} + k - 1))$

(b) For $k = 0$, when $f = \sigma \cdot 2^e$:

$f \approx x$ within 1 ulp of $x \Leftrightarrow f = x \cdot (1 + \varepsilon)$, $\varepsilon \in (-1 / (2^{N-1} + 1), 1 / (2^N - 1))$

Corollaries

Corollary 2

Let f and x as defined above, and m ϵ **R**, $0 < m \leq 1$

If $f \approx x$, within m ulps of x and m ϵ (0, 1/2) then $f = x \cdot (1 + \varepsilon)$, with $|\varepsilon| < m / (2^{N-1} + m)$

If $f \approx x$, within m ulps of x and m ϵ [1/2, 1], then $f = x \cdot (1 + \varepsilon)$, w/ $|\varepsilon| < m / (2^{N-1}+1-m)$

Corollary 3

Let f and x as defined above, and m ϵ **R**, $0 < m \leq 1$

If m ϵ (0, 1/2) and $f = x \cdot (1 + \varepsilon)$, with $|\varepsilon| < m / (2^{N} - m)$, then $f \approx x$, within m ulps of x

If m ϵ [1/2, 1] and $f = x \cdot (1 + \varepsilon)$, with $|\varepsilon| < m / (2^{N}+m-1)$, then $f \approx x$, within m ulps of x

Corollaries

Corollary 4

Let f and x as defined above

If $f \approx x$, within 1 ulp of x,
then $f = x \cdot (1 + \varepsilon)$, with $|\varepsilon| < 1/2^{N-1}$

Extension:

Let x, f ϵ **R***

If f is within 1 ulp of x on N bits,
then $f = x \cdot (1 + \varepsilon)$, with $|\varepsilon| < 1/2^{N-1}$

Corollary 5

Let f and x as defined above

If $f = x \cdot (1 + \varepsilon)$, with $|\varepsilon| < 1/2^{N}$,
then $f \approx x$, within 1 ulp of x

Extension:

Let x, f ϵ **R***

If $f = x \cdot (1 + \varepsilon)$, with $|\varepsilon| < 1/(2^{N} + 1)$,
then f is within 1 ulp of x on N bits

Corollaries

Corollary 6

Let f and x as defined above

If $f \approx x$, within 1/2 ulp of x, then $f = x \cdot (1 + \varepsilon)$, with $|\varepsilon| < 1/(2^N+1)$

To prove this property, substitute m = 1/2 in Corollary 2 above

Claude-Pierre Jeannerod (ENS) noticed that this recovers, as a special case, a result from Knuth's TAOCP (vol. 2, 3rd. ed., p. 232; 2nd ed. p. 216). This bound is attained for example at the midpoint $x = 1 + 2^{-N}$

Corollary 7

Let f and x as defined above

If $f = x \cdot (1 + \varepsilon)$, with $|\varepsilon| < 1/(2^{N+1}-1)$, then $f \approx x$, within 1/2 ulp of x

To prove this property, substitute m = 1/2 in Corollary 3 above

Existing literature provides similar properties for 1/2 ulp as those in Corollaries 6 & 7, however w/o '+1' and '-1'

Usage Example

Iterative Newton-Raphson Single Precision Division Algorithm, w/ CR Result

(1) $y_0 = 1/b \cdot (1 + \varepsilon_0)$, $|\varepsilon_0| \le 2^{-m}$, $m = 8.856$

(2) $e_0 = (1 - b \cdot y_0)_{rn}$ dbl.-ext. precision

(3) $y_1 = (y_0 + e_0 \cdot y_0)_{rn}$ dbl.-ext. precision

(4) $e_1 = (e_0^2)_{rn}$ dbl.-ext. precision

(5) $y_2 = (y_1 + e_1 \cdot y_1)_{rn}$ dbl.-ext. precision

(6) $q_0 = (a \cdot y_2)_{rn}$ single precision

(7) $r_0 = (a - b \cdot q_0)_{rn}$ single precision

(8) $q_1 = (q_0 + r_0 \cdot y_2)_{rnd}$ single precision

It was shown that $y_2 = 1/b \cdot (1 + err_2)$ with $|err| < 2^{-35.5439} < 1/(2^{35} + 1)$; from the extension to Corollary 5, y_2 is within 1 ulp of 1/b on 35 bits. It is thus (clearly) within 1/2 ulp of 1/b on N=24 bits (I)

This also means that q_0 before rounding satisfies $q_0^* = a/b \cdot (1 + err_2) < 1/(2^{35} + 1)$

Similarly, q_0^* is (clearly) within 1/2 ulp of 1/b on N=24 bits; rounding to nearest modifies it by at most 1/2 ulp, so $q_0 \approx a/b$ is within 1 ulp of a/b. (II)

With (I) and (II) true, a theorem can then be applied to (7) and (8), proving that q_1 is the correctly rounded value of a/b

Conclusion

- We established several simple, but useful relationships between ULP errors and the corresponding relative errors
- These can be used when converting between the two types of errors, ensuring that the least amount of information is lost in the process
- We used these in IEEE conformance proofs for iterative division and square root algorithms
- Could be useful to numerical analysts in:
  - 'hand' proofs of error bounds for floating-point computations
  - automated tools which carry out, or support deriving such proofs
- In some cases, we established tighter bounds than found in the literature
- We provided 'conversion' rules of finer granularity – also at FP value level, not only at binade level, and took into account the special conditions which occur at binade ends


intel®
experience
what's inside™