AN INTERLEAVED RATIONAL/RADIX ARITHMETIC SYSTEM FOR HIGH-PRECISION COMPUTATIONS

KAI HWANG AND T. P. CHANG

SCHOOL OF ELECTRICAL ENGINEERING, PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907

ABSTRACT - A new interleaved rational/radix number system is proposed for upgrading the precision of normalized Floating-Point (FLP) arithmetic operations without increasing the basic word length. A complete set of rational rounding and arithmetic algorithms are developed. The Average Relative Representation Error (ARRE) of the proposed flexible FLP system is computed through a series of simulation studies on CDC 6500. Our results show a 10% improvement of representation accuracy when compared with the ARRE of conventional FLP system. The architecture of a rational FLP arithmetic processor is also presented. Tradeoffs between operating speed and computing accuracy are discussed.

1. INTRODUCTION

This paper presents a new arithmetic system using Farey rationals interleaved with radix fractions for upgrading the mantissa precision of Floating-Point (FLP) arithmetic without increasing the basic word length. The precision of conventional FLP arithmetic is reflected essentially by the number, t, of significant bits in the fraction portion (mantissa) of FLP numbers. We used to consider each t-bit mantissa as a radix fraction with a fixed denominator 2^t and a numerator within the range $[0,2^t-1]$. Only the numerator needs to be represented in the radix number system, because the constant denominator is self-implied. The numerical gaps between adjacent radix fractions are uniform with a constant value of 2^{-t}. To simplify the subsequent discussions, we assume an even fraction length $t = 2k$ with $k \geq 1$ for all FLP numbers.

In Fig. 1, we propose a new interleaved fractional number system with two possible representations depending on the FLP numeric value being represented. A tag is used to distinguish the two fractional representations, radix fraction versus the Farey Rational to be described below. When tag = 1, a normalized 2k-bit radix fraction is represented with a leading bit 1 which coincides with the value of the tag bit. When tag = 0, an irreducible rational number $\frac{q}{p}$ within the range $\frac{1}{2} \leq \frac{q}{p} \leq 1$, called a Farey rational, is represented in its complemented form $\frac{p-q}{p}$. The left k-1 bit, hold complemented numerator p-q and the right k bits hold the denominator p. Note that $0 \leq \frac{p-q}{p} < \frac{1}{2}$ for normalized q. We shall formally define Farey rationals and describe their numerical properties in Section 2. Farey series has been discussed in number theory [5].

The use of Farey rationals is motivated from Matula's work [11]. Many real fractions, which can be written in our system as simple irreducible rationals as $\frac{1}{3}$, $\frac{5}{6}$, $\frac{4}{7}$, etc., can never be represented in closed form in conventional radix number system. Our analytic and simulation results show that with this "flexible" duel representations, the machine representation error can be significantly reduced. In other words, higher precision can be achieved in computer arithmetic with appropriate tagging and rounding schemes. We then compare our results with the precision performance of conventional FLP arithmetic systems as reported in Cody [2], Brent [1]

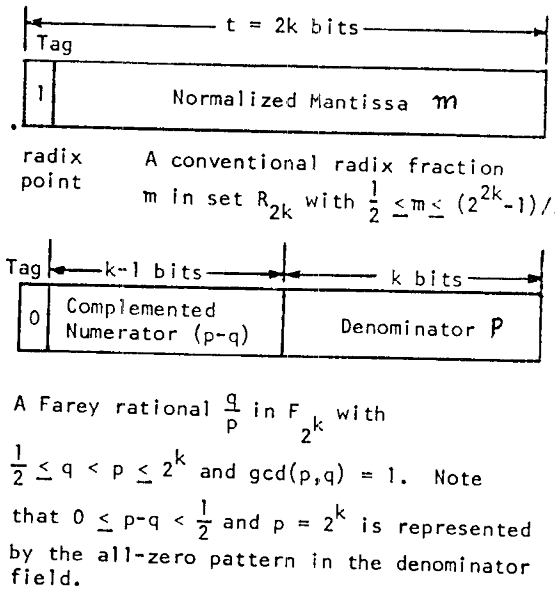


Fig. 1. The dual fraction representations: radix fraction vs. Farey rationals.

and Kuck et al [10], we shall consider base-2 normalized FLP numbers with fractions in the range $\left[\frac{1}{2}, 1\right]$. A comparison of our rational system with the p-adic [9] system for representing rationals is also given.

2. FAREY SERIES OF IRREDUCIBLE RATIONALS

A Farey series F_n of order n refers to the finite set of irreducible rationals within the closed interval [0, 1] whose denominators do not exceed n. Formally, we define

$$F_n = \{\frac{q}{p} \mid 0 \leq q < p \leq n \text{ and } \gcd(p,q) = 1\} \quad (1)$$

We shall consider Farey series of order $n = 2^k$, where $k = \log_2 n$ is the number of bits required to express the integer n in binary. Each member of F_n is called a Farey rational. The series of Farey rationals can be arranged in ascending order. Let $\frac{q}{p}$ and $\frac{v}{u}$ be any two Farey rationals in F_n such that $\frac{q}{p} < \frac{v}{u}$, then $\frac{q+v}{p+u}$ is defined as the mediant of $\frac{q}{p}$ and $\frac{v}{u}$. Described below are some useful properties associated with Farey rationals.

Theorem 1

If $\frac{q'}{p'} < \frac{q}{p} < \frac{q''}{p}$ are three consecutive Farey rationals in F_n, then

$$\frac{q}{p} = \frac{q' + q''}{p' + p''} \quad (2)$$

Proof of this mediant property can be found in Hardy and Wright [5]. One can repeatedly apply Theorem 1 to generate the entire Farey series F_n.

Algorithm for Generating Farey Rationals

Step 1 Start with $\frac{0}{1}$ and $\frac{1}{1}$ as the two extreme rationals and find their mediant $\frac{0+1}{1+1} = \frac{1}{2}$ as the first nontrivial Farey rational at the center of the series.

Step 2 Find the mediants of all existing pairs of Farey rationals until no more mediants with denominators $\leq$ n can be found.

The following properties are immediate from Theorem 1.

Corollary 1

Let $\frac{q'}{p'} < \frac{q}{p}$ be any two consecutive Farey rationals in F_n.

$$p'q - p\,q' = 1 \quad (3)$$

$$n + 1 \leq p + p' \leq 2n - 1 \quad (4)$$

The mediant $\frac{q+q'}{p+p'}$, falling in the interval $(\frac{q}{p}, \frac{q'}{p'})$, is not a member of F_n due to Property (4). The above corollary implies that no two consecutive Farey rationals can have the same denominator, as long as n > 1. Consider an example of n = 8 and k = 3. The Farey rationals in F_8 are listed below in ascending order.

$$F_8 = \{\frac{0}{1}, \frac{1}{8}, \frac{1}{7}, \frac{1}{6}, \frac{1}{5}, \frac{1}{4}, \frac{2}{7}, \frac{1}{3}, \frac{3}{8}, \frac{2}{5}, \frac{3}{7}, \frac{1}{2}, \frac{4}{7}, \frac{3}{5}, \frac{5}{8}, \frac{2}{3}, \frac{5}{7}, \frac{3}{4}, \frac{4}{5}, \frac{5}{6}, \frac{6}{7}, \frac{7}{8}, \frac{1}{1}, \} \quad (5)$$

Corollary 2

Given two integers i < j, then all fractions with denominator i, $\frac{h}{i}$ for h = 1, 2, 3,..., j-1, can be reduced to be irreducible rationals in F_j.

This corollary shows that we can generate the entire Farey series F_n by simply listing all the rationals with denominators increasing from 2 to n. Repeated appearance of rationals with the same irreducible value should be crossed out from the list. For n = 8, the Farey series F_8 can be generated by the following triangular listing of all irreducible rationals with denominators strictly less than 9.

$\frac{1}{2}$

$\frac{1}{3}, \frac{2}{3}$

$\frac{1}{4}, \cancel{\frac{2}{4}}, \frac{3}{4}$

$\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}$

$\frac{1}{6}, \cancel{\frac{2}{6}}, \cancel{\frac{3}{6}}, \cancel{\frac{4}{6}}, \frac{5}{6}$

$\frac{1}{7}, \frac{2}{7}, \frac{3}{7}, \frac{4}{7}, \frac{5}{7}, \frac{6}{7}$

$\frac{1}{8}, \cancel{\frac{2}{8}}, \frac{3}{8}, \cancel{\frac{4}{8}}, \frac{5}{8}, \cancel{\frac{6}{8}}, \frac{7}{8},$

plus $\frac{0}{1}$ and $\frac{1}{1}$

The gaps between adjacent Farey rationals are not uniform. We shall demonstrate the gap distributions of Farey series in section 4. The following theorem provides some bounds on these nonuniform gaps.

Theorem 2

Let g be the gap between any two adjacent Farey rationals in F_n. Then

$$\frac{1}{n(n-1)} \leq g \leq \frac{1}{n} \quad (6)$$

Proof:

Consider two adjacent Farey rational $\frac{q}{p} < \frac{q'}{p'}$ in F_n.

$$q = \frac{q'}{p'} - \frac{q}{p} = \frac{q'p - p'q}{pp'}$$

By Corollary 1, $pq' - p'q = 1$, we have $g = \frac{1}{pp'}$. By the fact that $1 \leq p \leq 2^k$, $1 \leq p' \leq 2^k$, $p \neq p'$, and $n + 1 \leq p + p' \leq 2n-1$, we have

$Max(pp') = n \cdot (n-1),$

$Min(pp') = 1 \cdot n$

Therefore (6) is proved by the following inequality

$$\frac{1}{Max(pp')} \leq g \leq \frac{1}{Min(pp')} \qquad \text{Q.E.D.}$$

3. THE UNION SPACE OF RADIX FRACTIONS AND FAREY RATIONALS

Let $R(0,1) = [0, 1]$ be the set of real numbers within the unity interval and R_t be the set of t-bit radix fractions. Of course $R_t \subset R(0,1)$ for all integers $t > 0$. A union space U_t is defined as the union of the set of radix fractions R_t and Farey series $F_{2^{t/2}}$ of order $2^{t/2}$. i.e.

$$U_t = R_t \cup F_{2^{t/2}} \tag{7}$$

With $t = 2k$, we write

$$U_{2k} = R_{2k} \cup F_{2^k} \tag{8}$$

The union space U_{2t} is formed by inserting the Farey rationals of F_{2^k} into the uniform gaps of the fraction set R_{2k}. The distribution of these interleaved fractions in the union space U_{2k} is symmetrical with respect to the center point of 1/2. The dual fraction representations described in Fig. 1 correspond to all normalized fractions in the upper half of the union space $U_t = U_{2k}$. Normalized FLP arithmetic operations over this union space will be defined in section 5.

Theorem 3

The intersection of the fraction set R_{2k} with *the Farey series F_{2^k} equals the fraction set R_k*

$$R_{2k} \cap F_{2^k} = R_k \tag{9}$$

Proof:

By definition, $R_k \subset F_{2^k}$ and $R_k \subset R_{2k}$

Thus, we have $R_k \subset F_{2^k} \cap R_{2k}$ (a)

Now consider any $\frac{q}{p} \in R_{2k} \cap F_{2^k}$.

Then $\frac{q}{p} = \frac{1 \cdot m}{p \cdot m} = \frac{1 \cdot m}{2^{2k}} \in R_{2k}$ for some m. This means that p divides 2^{2k} or, in turn, p divides 2^k. Therefore, there exists an h such that

$$\frac{q}{p} = \frac{q \cdot h}{p \cdot h} = \frac{q \cdot h}{2^k} \in R_k$$

Therefore, we have $F_{2^k} \cap R_{2k} \subset R_k$ (b)

(a) and (b) completes the proof.

Q.E.D.

Theorem 4

Let $t = 2k$ and $a = \frac{w}{2^t}$ and $b = \frac{w+1}{2^t}$ be two adjacent fractions in set R_t for some $0 \leq w \leq 2^t - 1$. There can be at most one Farey rational $\frac{q}{p} \in F_{2^k}$ in the interval [a,b], i.e.

$$\frac{w}{2^t} \leq \frac{q}{p} \leq \frac{w+1}{2^t} \tag{10}$$

Proof: Consider any two adjacent Farey rationals $\frac{q}{p} < \frac{q'}{p'}$ in F_{2^k}. By Theorem 2, we know that $\frac{q}{p}$ and $\frac{q'}{p'}$ must be at least $\frac{1}{2^k(2^k-1)}$ distance apart, which is longer than the uniform gap 2^{-2k} between adjacent fractions in R_{2k}. Therefore, in each gap of R_t, there exists either one or none Farey rational from F_{2^k}.

Q.E.D.

Figure 2 shows the set-theoretic relationships among a number of subsets of fractions. These fraction subsets will be used to specify various arithmetic functions over the Cartisian product space $U_{2k} \times U_{2k}$ of the union space. The union space corresponds to shaded area in Fig. 2.

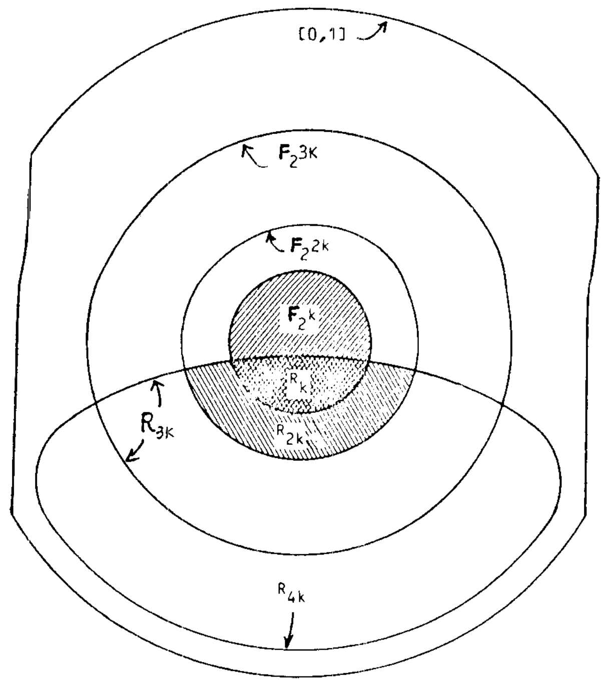


Fig. 2. Set - theoretic relationships among a number of subsets of fractions in the unit interval [0, 1].

The gaps between adjacent Farey rationals in F_{2^k} are not uniformly distributed. In fact, it assumes the symmetric distribution pattern as demonstrated in Fig. 3 for F_{32}. The curve is symmetrical with respect to the rational 1/2 at the center. Two largest gaps with value 2^{-k} occur between $(\frac{0}{1}, \frac{1}{2^k})$ and between $(\frac{2^k-1}{2^k}, \frac{1}{1})$. The smallest gap equals $\frac{1}{2^k \cdot (2^k-1)}$, which is about 2^k times smaller than the maximum gap for larger k. The average gap in F_{32} are about ten times smaller than the maximum gap in F_{32}. When the word length increases, the average gap tends to decrease rapidly.

The gap probability distributions associated with four Farey series with increasing orders of 16, 32, 128, and 256 are demonstrated in Fig. 4. As the word length increases, the gap distribution tends to become a delta function near the zero. This means that most gaps are small when k is sufficiently large. Only a handful of gaps appear as spikes with very low probability. The gap distribution for the union space U_8 is shown in Fig. 5. Most gaps in U_{2k} assume the value 2^{-2k} as shown by the flat peaks at the top of the drawing. Small gaps between Farey rationals and radix fractions appear as steep ditches.

4. RATIONAL ROUNDING SCHEMES

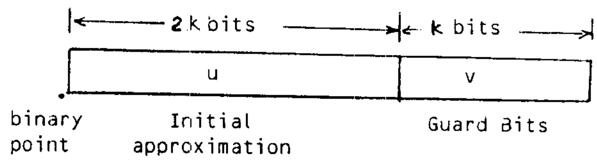
Consider an arbitrary real fraction $x \in R$, where $R = [\frac{1}{2}, 1]$ is the upper half of the unity interval. This real number x can be approximated by a machine number in space U_{2k} in either radix form or rational form through the following rounding operations. First, we retain the leading 3k bits of x through an α-mapping

$$\alpha : R \rightarrow R_{3k} \tag{11}$$

Then, we apply a rounding transformation to produce the machine representation

$$\rho : R_{3k} \rightarrow U_{2k} \tag{12}$$

This ρ-mapping maps every 3k-bit fraction in R_{3k} into a 2k-bit number in U_t, which appears either as a 2k-bit normalized fraction $m \in R_{2k}$ or a normalized Farey rational $\frac{q}{p} \in F_{2^k}$, depending on which of these two representations results in less error. The fraction $y=\alpha(x)$ can be written as the sum of two subtractions as shown below:



$$y = u \cdot 2^{-2k} + v \cdot (2^{-2k} - 2^{-3k}), \tag{13}$$

where u and v are 2k-bit and k-bit integers. Obviously, the subfraction $u \cdot 2^{-2k}$ can be used as an initial approximation of $z=\rho(y)$.

Assume $y \in [a,b]$, where $a = w/2^{2k}$ and $b = (w+1)/2^{2k}$ are two adjacent fractions in R_{2k} There are two cases to be considered in realizing the rounding operation ρ in the union space U_{2k}.

Case 1. No Farey rational lies in [a,b]. The value of $\rho(y)$ is determined by the nearest neighborhood rounding as usual. That is

$$\rho(y) = \begin{cases} a, & \text{if } a \leq y \leq \frac{a+b}{2} \\ b, & \text{if } \frac{a+b}{2} < y \leq b \end{cases} \tag{14}$$

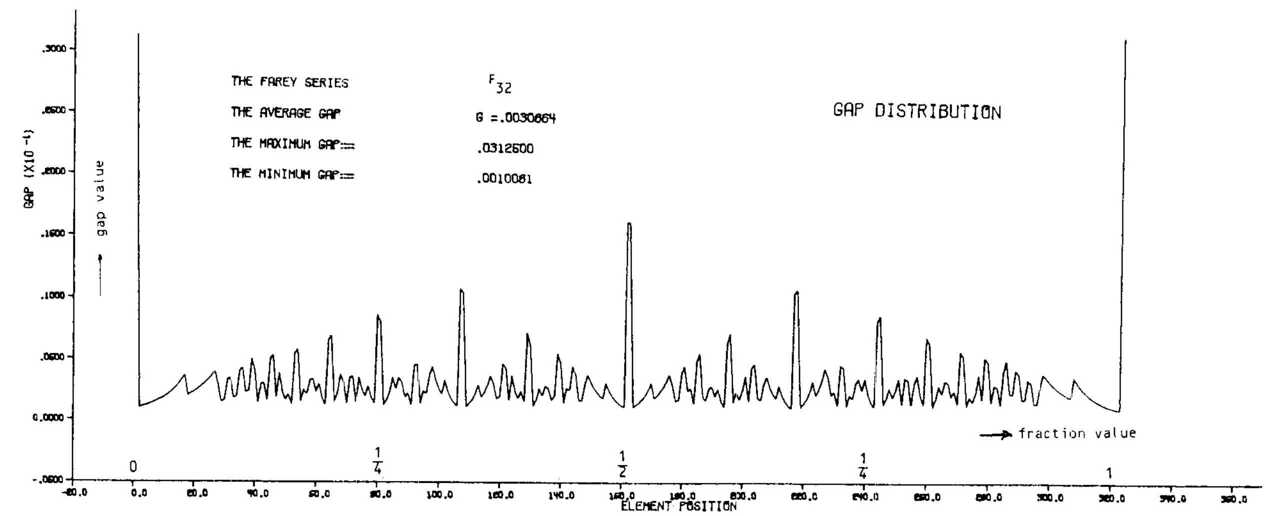


Fig. 3. The gap distribution of Farey Series F_{32} of order 32.

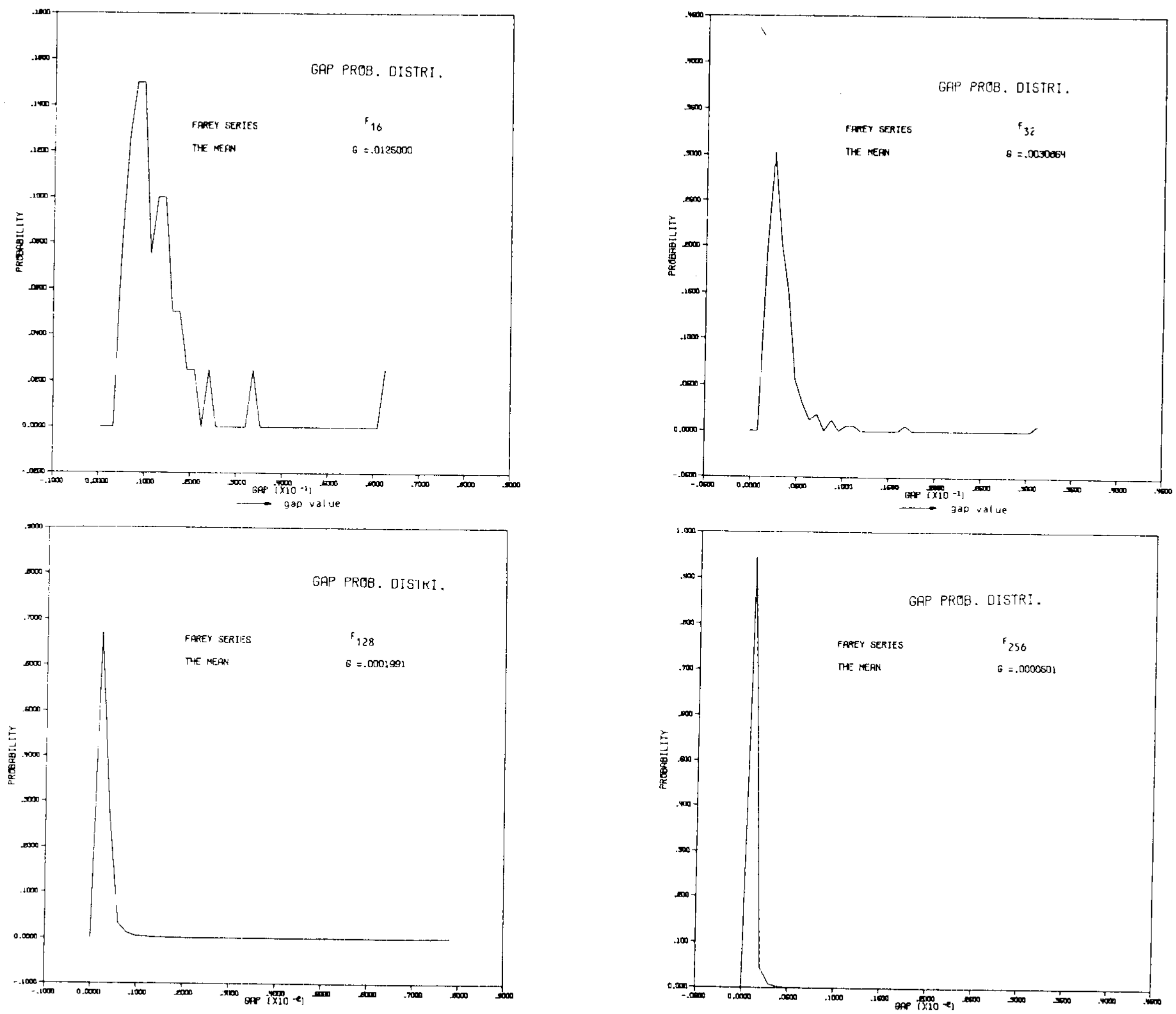


Fig. 4. The probability distributions for Farey Series F_{16}, F_{32}, F_{128}, and F_{256} respectively.

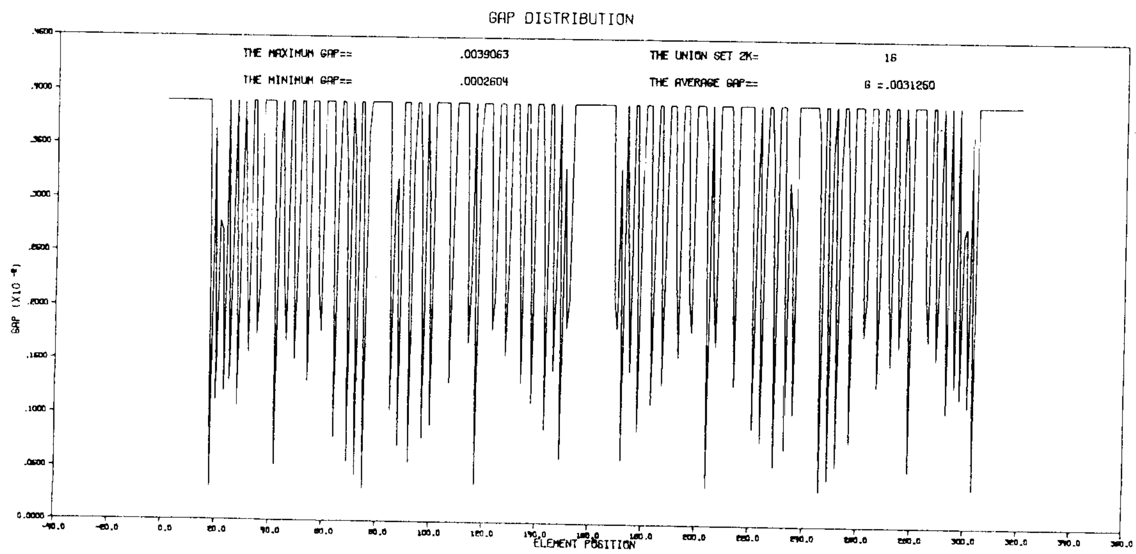


Fig. 5. The gap distribution of the union space U_{16}.

Case 2. There is a Farey rational f [a,b] such that $a < f < b$. Assume $y \in [a,f]$. (The case of $y \in [f,b]$ can be similarly discussed).

$$\rho(y) = \begin{cases} a, & \text{if } |y-a| < |y-f| \\ f, & \text{if } |y-a| \geq |y-f|, \end{cases} \tag{15}$$

where $|y-a|$ and $|y-f|$ are the absolute distance between y and points a and f respectively.

In both cases, we need to first decide whether there is a Farey rational, f, lying in the interval [a,b]. This can be done by the following recursive procedures in finding the Euclidean convergence to points a and b respectively.
We can write the point

$$a = \frac{w}{2^{2k}} = \frac{U_0}{V_0} = [a_0, a_1, a_2, \cdots, a_d]$$

$$= a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots \, a_{d-1} + \cfrac{1}{\cfrac{1}{a_d}}}}} \tag{16}$$

where integer $a_i = [\frac{U_i}{V_i}]$ for $i \geq 0$, and for $i \geq 1$, we have

$$U_i = U_{i-1} - a_{i-1} \cdot V_{i-1}$$
$$V_i = U_{i-1} \tag{17}$$

We define $C_i = \frac{q_i}{p_i}$ as the successive convergence to a by the following procedure with initial conditions $C_0 = q_0 = 0$ and $p_0 = p_1 = q_1 = 1$ and for $i \geq 2$.

$$q_i = a_i \cdot q_{i-1} + q_{i-2}$$
$$p_i = a_i \cdot p_{i-1} + p_{i-2} \tag{18}$$

Find the smallest i, say $i = d$, such that $p_d > 2^k$, then the nearest Farey fraction to point a is obtained as $f = C_{d-1} = \frac{q_{d-1}}{p_{d-1}}$. It is interesting to point out that $f \leq a$ when d is even, and that $f > a$ when d is odd. Similarly, we can decide whether $f \leq b$ or $f > b$. The two margin detections reveal the fact whether there is a f in [a,b] or not. The nearest neighborhood test is then applied to obtain the final value $z = \rho(y) = \rho(\alpha(x)) = \rho \cdot \alpha(x)$. To do this, the use of the right k guard digits guarantee that minimum gaps are used to ensure the optimality of the rounding scheme in U_{2k}.

Detailed descriptions of above rounding procedures written in Pidgin ALGOL are given in the full-length paper [7].

The following result proved in [5] can be used to estimate the rounding errors associated with the proposed system.

Theorem 5

If x is an arbitrary real fraction in R(0,1), then there exists a Farey rational $\frac{v}{u} \in F_2k$ such that

$$\frac{1}{u \cdot (2^{k+1}+1)} \leq |x - \frac{v}{u}| \leq \frac{1}{u \cdot (2^k+1)} \tag{19}$$

For large k, this bound can be written as $\frac{2^{-(k+1)}}{u} \leq |x - \frac{v}{u}| \leq \frac{2^{-k}}{u}$.

5. RATIONAL ARITHMETIC ALGORITHMS

Four rational arithmetic operations, namely RADD, RSUB, RMPY and RDIV, are to be defined below in terms of composite mappings from set $U_{2k} \times U_{2k}$ to set U_{2k}. The operand pairs in set $U_{2k} \times U_{2k}$ can be divided into four subspaces

$$U_{2k} \times U_{2k} = (R_{2k} \cup F_2k) \times (R_{2k} \cup F_2k)$$
$$= R_{2k} \times R_{2k} \cup R_{2k} \times F_2k \cup F_2k \times R_{2k}$$
$$\cup F_2k \times F_2k \tag{20}$$

It suffices to consider arithmetic operations on three subspaces $R_{2k} \times R_{2k}$, $R_{2k} \times F_2k$, and $F_2k \times F_2k$. Operations defined on subspace $F_2k \times R_{2k}$ are similar to those for $R_{2k} \times F_2k$. Only normalized radix fractions $\frac{1}{2} \leq \frac{m}{2^{2k}} < 1$ from R_{2k} and normalized Farey rationals $\frac{1}{2} \leq \frac{q}{p} < 1$ from F_2k are considered legitimate operands. Operands from set F_2k represented in complemented from, $\frac{p-q}{p}$, must be converted to normal form, $\frac{q}{p}$, before they can be applied in rational arithmetic operations.

Four standard fixed-point arithmetic operations, denoted as $\oplus$, $\ominus$, $\otimes$, and $\oslash$, and four auxiliary operations denoted as α, β, δ, and ρ, used to define rational arithmetic operations associated with mantissa arithmetic in FLP processors. The mappings α and ρ were defined in (11) and (12) respectively. β is a left-shift operation which shifts a radix fraction or the numerator of a Farey rational one bit to the left. δ can be similarly defined except shifting to the right. β is needed for normalization purpose. δ is needed for operand alignment to avoid mantissa sum overflow or quotient overflow. We shall first define RMPY and RDIV operations.

Rational Multiplication (RMPY)

The product of two normalized fractions must be in $[\frac{1}{4},1)$. Normalization is required only when the product is in $[\frac{1}{4},\frac{1}{2})$.

RMPY:

$$
\begin{aligned}
&R_{2k} \times R_{2k} \xrightarrow{\otimes} R_{4k} \xrightarrow{\beta} R_{4k} \xrightarrow{\alpha} R_{3k} \xrightarrow{\rho} U_{2k},\\
&R_{2k} \times F_2k \xrightarrow{\otimes} F_23k \xrightarrow{\beta} F_23k \xrightarrow{\alpha} R_{3k} \xrightarrow{\rho} U_{2k},\\
&F_2k \times F_2k \xrightarrow{\otimes} F_22k \xrightarrow{\beta} F_22k \xrightarrow{\alpha} R_{3k} \xrightarrow{\rho} U_{2k}.
\end{aligned}
\tag{23}
$$

By considering $R_{2k} \subset F_22k$, the operations $R_{2k} \times F_2k \xrightarrow{\otimes} F_23k$ and $F_{2k} \times F_2k \xrightarrow{\otimes} F_22k$ are performed by multiplying the corresponding numerators and denominators separately. The β-operation may be skipped if the initial produce after $\otimes$ is already normalized in $[\frac{1}{2},1)$.

Rational Division (RDIV)

Due to the fact that $\frac{b}{a}/\frac{d}{c} = \frac{b \cdot c}{a \cdot d}$, RDIV can be defined similarly to RMPY. If the dividend is greater than the divisor, one left shift is needed on the dividend to avoid quotient overflow. No normalization is needed, when both operands are normalized and properly aligned.

RDIV:

$$
\begin{aligned}
&R_{2k} \times R_{2k} \xrightarrow{\delta} R_{2k} \times R_{2k} \xrightarrow{\otimes} F_22k \xrightarrow{\alpha} R_{3k} \xrightarrow{\rho} U_{2k}\\
&R_{2k} \times F_{2k} \xrightarrow{\delta} R_{2k} \times F_{2k} \xrightarrow{\otimes} F_23k \xrightarrow{\alpha} R_{3k} \xrightarrow{\rho} U_{2k}\\
&F_{2k} \times F_2k \xrightarrow{\delta} F_2k \times F_2k \xrightarrow{\otimes} F_22k \xrightarrow{\alpha} R_{3k} \xrightarrow{\rho} U_{2k}
\end{aligned}
\tag{24}
$$

In FLP arithmetic, exponents must be equalized before mantissa addition or subtraction are performed. One can increment the smaller exponent to match the larger one and at the same time shift right the mantissa associated with the smaller exponent. Suppose we start with two normalized FLP numbers. After equalizing the exponents, we may end up with one normalized mantissa and one unnormalized mantissa. The sum of these two mantissas lies in range $[\frac{1}{2},2)$. We shall denote

$$R^{*}_{2k} = \{x \mid x \in [\tfrac{1}{2},2) \text{ and } \tfrac{x}{2} \in R_{2k}\}$$

$$\text{and } F^{*}_{2k} = \{\tfrac{q}{p} \mid \tfrac{q}{p} \in [\tfrac{1}{2},2) \text{ and } \tfrac{q/2}{p} \in F_2k\}$$

The division-by-2 operation corresponds to right-shift operation δ.

Rational Addition (RADD)

$$
\begin{aligned}
&R_{2k} \times R_{2k} \xrightarrow{\oplus} R^{*}_{2k} \xrightarrow{\delta} R_{2k+1} \xrightarrow{\rho} U_{2k}\\
&R_{2k} \times F_{2k} \xrightarrow{\oplus} F^{*}_23k \xrightarrow{\delta} F_23k \xrightarrow{\alpha} R_{3k} \xrightarrow{\rho} U_{2k}\\
&F_2k \times F_2k \xrightarrow{\oplus} F^{*}_22k \xrightarrow{\delta} F_22k \xrightarrow{\alpha} R_{3k} \xrightarrow{\rho} U_{2k}
\end{aligned}
\tag{25}
$$

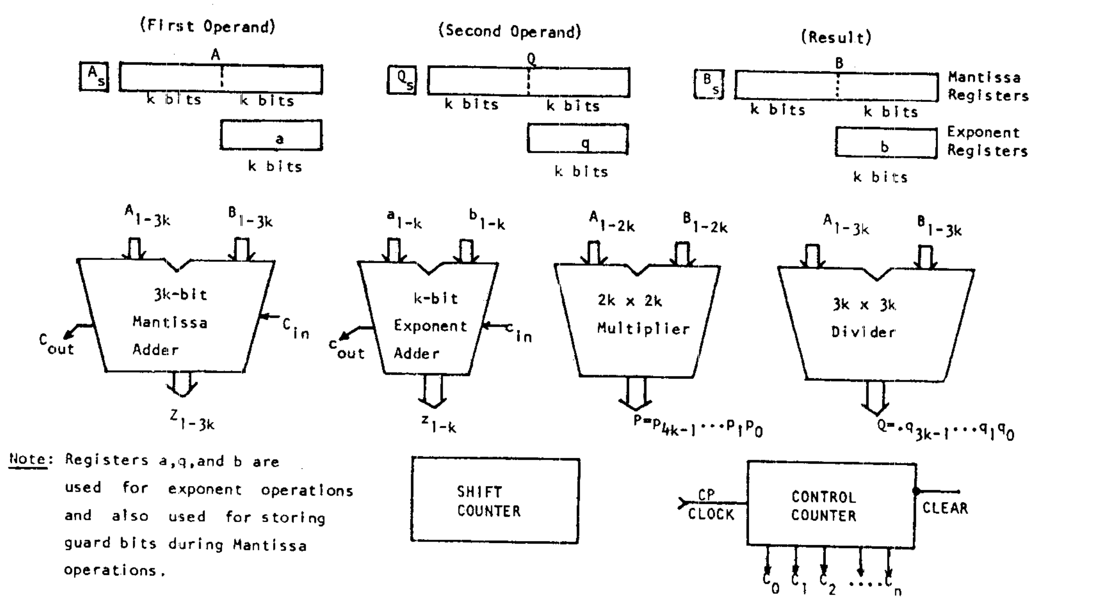


Fig. 6. Basic hardware components in a rational arithmetic processor for normalized FLP arithmetic with mantissa in interleaved radix/rational form.

Rational Subtraction (RSUB) can be defined similarly to RADD. Detailed procedures in flowchart form for the above rational operations have been described in [7]. The architecture of a FLP arithmetic processor for implementing the above rational mantissa operations is proposed in Fig. 6. Three sets of registers are used. A, B, and Q are mantissa registers (each 2k-bits) and a, b, and q are exponent registers (each k-bits). In addition to one Mantissa Adder and one Exponent Adder, the system requires also a hardware Multiplier and a divider as shown.

6. ARRE OF RATIONAL ARITHMETIC

Mckeeman [13] and Cody [2] have derived analytic formulas for measuring the Average Relative Representation Errors (ARRE) associated with conventional FLP arithmetic systems with different word lengths and base values. Recently Kuck [10] and his associates provided a comparative study of ARRE's caused by various rounding schemes. In this section, we present an ARRE evaluation of the proposed dual-representation system for mantissas of FLP numbers. Our results are then compared with those associated with conventional systems. We consider normalized FLP system with a logarithmic probability distribution [4] for all fractions between $\frac{1}{r} \leq x < 1$, where r is the base value.

$$P(x) = \frac{1}{x \cdot \ln r} \tag{26}$$

The relative representation error associated with approximating an arbitrary real fraction x by a t-bit machine fraction is defined as

$$|Q(x)| = \left|\frac{\rho(x) - x}{x}\right| \tag{27}$$

where $\rho(x)$ is the machine number obtained by applying an appropriate rounding scheme ρ. In general, the ARRE can be defined to be

$$\mathrm{ARRE} = \int_{\frac{1}{r}}^{1} |Q(x)| \cdot P(x)\,dx \tag{28}$$

In the conventional FLP system with the set R_t of uniformly distributed radix fraction of t bits long, the error function Q(x) can be approximated by

$$Q(x) = \frac{\underset{\frac{1}{r} \leq x < 1}{\mathrm{Ave}} |\rho(x) - x|}{x} = \frac{2^{-t}/4}{x} \tag{29}$$

Therefore, for all $\rho(x) \in R_t$, we obtain the following closed result

$$\mathrm{ARRE}(R_t) = \int_{\frac{1}{r}}^{1} \frac{1}{x \cdot \ln r} \cdot \frac{2^{-t}}{4x}\,dx = 2^{-(t+2)}/\ln r \tag{30}$$

For our dual representation system, the gap distribution between adjacent elements in U_t is not uniform. The error function Q(x) cannot be written in closed form for all $\rho(x) \in U_t$. We can still evaluate the ARRE for the union space U_t by computing the following summation series with uniform increment Δ smaller than the minimum gap in U_t. Let $x_o = \frac{1}{r}$ and integer $j = \left\lceil \frac{1 - 1/r}{\Delta} \right\rceil$. We are considering j equally spaced fractions in interval $\left[\frac{1}{2}, 2\right]$ as sample points in our simulation study.

$$x_i = x_o + i \cdot \Delta = \frac{1}{r} + i\Delta \text{ for } i = 0,1,2,\ldots, j-1 \tag{31}$$

We evaluate Eq. 28 for all $\rho(x) \in U_t$ by

$$\begin{aligned} \mathrm{ARRE}(U_t) &= \sum_{i=0}^{j-1} P(x_i) \cdot Q(x_i) \cdot \Delta \\ &= \sum_{i=0}^{j-1} \frac{\Delta}{x_i \ln r} \left|\frac{\rho_u(x_i) - x_i}{x_i}\right| \\ &= \frac{1}{\ln r} \sum_{i=0}^{j-1} \frac{\Delta \cdot \left|\rho_u(\frac{1}{r} + i\Delta) - (\frac{1}{r} + i\Delta)\right|}{(\frac{1}{r} + i\Delta)^2} \end{aligned} \tag{32}$$

where ρ_u: $\{x_i | 0 \leq i \leq j-1\} \to U_t$ is the rounding method specified in Eqs. 11 and 12.

Table 1 shows a comparison of the gap characteristics and the ARRE's associated with three fraction number systems, R_{2k}, F_{2k}, and U_{2k}, for base r = 2 and two fraction word lengths 2k = 8 and 16. The increment Δ used in each simulation experiment was chosen to be a fraction of the minimum gap in each case.

The $\mathrm{ARRE}(U_{2k})$ associated with different k have been computed by simulation experiments on the CDC 6500 computer at Purdue University. A comparison of ARRE (R_t) and ARRE (U_t) is given in Table 2 and plotted in Fig. 7 for base r = 2 and word length from 6 to 20 bits. The increment used for the union space U_{20} is equal to $\Delta = 0.240 \times 10^{-7}$. It took 4.4 hours CPU time of CDC 6500 to compute the value of ARRE (U_{20}).

An ARRE Improvement Factor $\theta(2k)$ is defined below to compare the relative performance of ARRE (U_{2k}) over the conventional ARRE (R_{2k}).

$$\theta(2k) = \frac{\mathrm{ARRE}\ (R_{2k}) - \mathrm{ARRE}\ (U_{2k})}{\mathrm{ARRE}\ (R_{2k})} \tag{33}$$

The $\theta(2k)$ is plotted in Fig. 8 for word lengths from 6 to 20 bits. The improvement factor $\theta(2k)$ tends to fluctuate around 10% for all word lengths greater than 8 bits. This means that our proposed rational arithmetic system is always 10% better in precision than the conventional radix

Table 1. Computer Simulation Results of Numerical Characteristics of Various Number Systems.

property / system	Minimum Gap		Maximum Gap		Average Gap		ARRE	
	2k=8	2k=16	2k=8	2k=16	2k=8	2k=16	2k=8	2k=16
Radix Fractions R_{2k}	0.391×10^{-2}	0.153×10^{-4}	0.391×10^{-4}	0.153×10^{-4}	0.391×10^{-2}	0.153×10^{-4}	0.141×10^{-4}	0.550×10^{-7}
Farey Rationals F_{2^k}	0.417×10^{-2}	0.153×10^{-4}	0.625×10^{-1}	0.391×10^{-2}	0.125×10^{-1}	0.501×10^{-4}	0.956×10^{-2}	0.750×10^{-4}
Union Set U_{2k}	0.260×10^{-3}	0.598×10^{-7}	0.391×10^{-2}	0.153×10^{-4}	0.313×10^{-2}	0.117×10^{-4}	0.125×10^{-4}	0.495×10^{-7}

Table 2. Comparison of ARRE (R_{2k}) and ARRE (U_{2k}) for Various Word Lengths.

Word Length	6	8	10	12	14	16	18	20
ARRE (R_{2k})	5.64×10^{-3}	1.41×10^{-3}	3.52×10^{-4}	8.81×10^{-5}	2.20×10^{-5}	5.50×10^{-6}	1.38×10^{-6}	3.44×10^{-7}
ARRE (U_{2k})	4.39×10^{-3}	1.25×10^{-3}	3.14×10^{-4}	7.91×10^{-5}	1.98×10^{-5}	4.95×10^{-6}	1.23×10^{-6}	3.09×10^{-7}

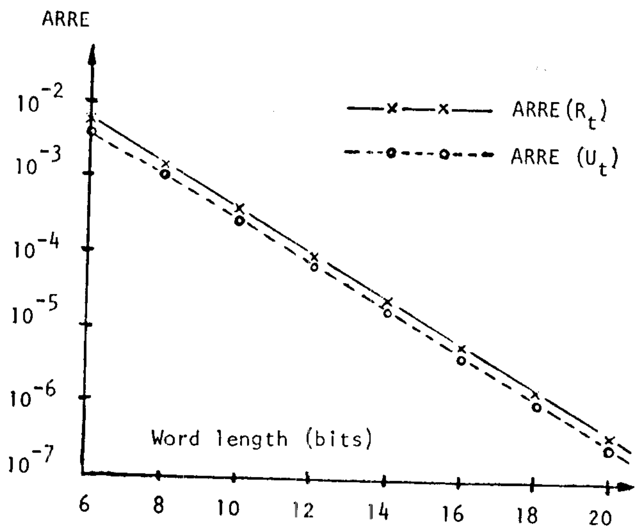


Fig. 7. ARRE's for radix fractions in R_t and for union space U_t versus various fraction length.

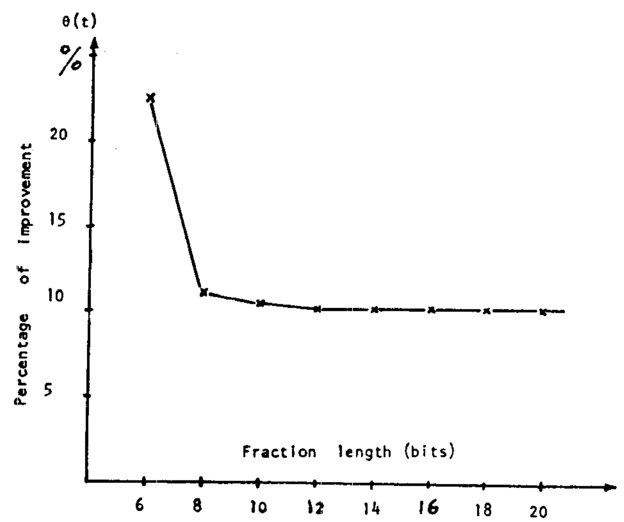


Fig. 8. The ARRE improvement factor $\theta(t)$ versus word length t.

system. This is considered a significant improvement over the conventional approach due to the accumulative nature of representational errors in a sequence of numerical computations.

7. CONCLUSIONS

The proposed FLP arithmetic system using interleaved Farey rationals and radix fractions results in significant (10%) increase in representation accuracy of machine arithmetic without extending the operand word length. Number-theoretic analysis and extensive computer simulation results are reported. The numerical simulation experiments verify the theoretical results nicely. Our flexible dual-representation system complements the p-adic rational arithmetic developed by Krishnamurthy [9] in the sense that our system can be immediately applied to existing FLP arithmetic computations in general. Of course, there exists the tradeoff between representation accuracy and computation speed. We gain the accuracy at the expense of increased computation overhead. However, with pipelined design of the proposed rational/radix arithmetic processor, the problem of increased delay due to computation overhead can be greatly alleviated. Continued efforts should be conducted in developing such high-speed pipelined rational arithmetic processors.

REFERENCES

[1] Brent, R. P., "On the Precision Attainable with Various Floating-point Number System," IEEE Trans. Comput., Vol. C-22, June 1973, pp. 601-607.

[2] Cody, W. J., "Static and Dynamic Numerical Characteristics of Floating Point Arithmetic," IEEE Trans. Comput., Vol. C-22, June 1973, pp. 598-601.

[3] Grosswald, E., Topics from the Theory of Numbers, The Macmilliam Co., New York, 1966.

[4] Hamming, R. W., "On the Distribution of Numbers," Bell Syst. Tech. J., Vol. 49, Oct. 1970, pp. 1609-1626.

[5] Hardy, G. H. and Wright, E. M., An Introduction to the Theory of Numbers, Clarendon Press, Oxford, 1960, 23-37, pp. 128-152.

[6] Hwang, K., Computer Arithmetic: Principles, Architecture, and Design, John Wiley Inc., New York, 1978.

[7] Hwang, K. and Chang, T. P., "High-Precision Floating-Point Arithmetic with Interleaved Farey Rationals and Radix Fractions," (to appear).

[8] Knuth, D. E., The Art of Computer Programming, Vol. 2, Semiannual Algorithms, Addison-Wesley, Menlo Park, CA, 1969.

[9] Krishnamurthy, E. V., "Matrix Processors Using P-adic Arithmetic for Exact Linear Computations," IEEE Trans. Comp., July 1977, pp. 633-639.

[10] Kuck, D. et al., "Analysis of Rounding Methods in Floating-Point Arithmetic, IEEE Trans. Computers, Vol. C-26, No. 7, July 1977, pp. 643-650.

[11] Matula, D. W., "Number Theoretic Foundation of Finite Precision Arithmetic," in Applications of Number Theory to Numerical Analysis, (W. Zaremder, ed.), Academic Press, New York, 1972, pp. 479-489.

[12] Matula, D. W., "Fixed-slash and Floating-slash Rational Arithmetic," in Proc. 3rd IEEE Symp. on Computer Arithmetic, Dallas, TX, Nov. 1975, pp. 90-91.

[13] McKeeman, W. M. "Representation Error for Real Numbers in Binary Computer Arithmetic," IEEE Trans. Electron Comput., Vol. EC-16, Oct. 1967, pp. 682-683.

[14] Yohe, J. Michael, "Rounding in Floating-point Arithmetic," IEEE Trans. on Computers, Vol. C23, No. 6, June 1973, pp. 577-586.