

either by a ring counter or a shift register with a *one* rippling through.

VI. CONCLUSION

A BIN/BCD conversion method has been developed which lends itself to unlimited expansion by using the geometrical similarity of interconnecting maps. Properly designed, the maps then contain all the necessary information to determine the size, the content, and the actual wiring of the decoding ROM's. Wiring diagrams were used to develop the hybrid conversion scheme. Conversion systems for practically any speed or size can be designed by using either the static or the hybrid method. This scheme is not limited to binary/BCD conversion, but can be extended to other types of conversion as well, such as, synchro/BCD, etc.

REFERENCES

- [1] J. F. Couleur, "A binary to decimal or decimal to binary converter," *IRE Trans. Electron. Comput.*, vol. EC-7, p. 313, 1958.
- [2] Z. M. Benedek and B. Moskowitz, "Convert binary to BCD without flip-flops," *Electron. Des.*, p. 58, October 10, 1968.
- [3] J. Linford, "Code conversion with semiconductor read only memories," Motorola Application Note 506.

On the Use of Continued Fractions for Digital Computer Arithmetic

KISHOR S. TRIVEDI

Abstract—Recently, there has been some interest in the use of continued fractions for digital hardware calculations. We require that the coefficients of the continued fractions be integral powers of 2 and, therefore, well-known continued fraction expansions of functions cannot be used. Methods of expansion of a large number of functions are presented. We show that the problem of selection of coefficients of the continued fractions does not have practical solution in most of the cases we have considered.

Index Terms—Bilinear transformation, computer arithmetic, continued fractions, hardware, logarithm, number system, quadratic equation, redundancy, Riccati equation, selection procedure.

I. INTRODUCTION

In this study, we have investigated the possibility of using continued fractions to evaluate elementary functions in hardware. A continued fraction is represented by

$$\frac{p_1 p_2}{q_1 + q_2} + \dots,$$

where p_i is known as a partial numerator and q_i is known as a partial denominator. An n -term approximation to such a con-

tinued fraction, denoted by P_n/Q_n , can be obtained using the following recursions [1]:

$$\left. \begin{aligned} P_0 &= Q_{-1} = 0, Q_0 = P_{-1} = 1 \\ P_{i+1} &= p_{i+1}P_{i-1} + q_{i+1}P_i \\ Q_{i+1} &= p_{i+1}Q_{i-1} + q_{i+1}Q_i \end{aligned} \right\} \quad (1.1)$$

In order to reduce the four multiplications in the above recursions to shifts, we require that the partial numerators and denominators be integral powers of 2. As a result of this restriction we are not able to use well-known continued fraction expansions of the functions to be evaluated. For example, to evaluate $\tanh x$, we may not use the expansion:

$$\tanh x = \frac{xx^2}{1+3} + \dots + \frac{x^2}{2n+1} + \dots$$

The first step in this direction was taken by deriving a method of expansion for the solution to a quadratic equation [2]. The class of Riccati differential equations is closed under a bilinear transformation [3]. In this correspondence we show that using this approach, a large number of elementary functions can be expanded into a continued fraction. We also present a new method of expanding $\log_e x$ into a continued fraction.

We would like to keep the set of allowable values of the partial numerators and denominators small. Once these two sets of allowable values are chosen, the range of numbers representable as continued fractions is fixed and finite. This introduces a restriction to the possible values of p_i and q_i at an iterative step. Furthermore, since the value of the function to be evaluated is known only implicitly through some coefficients, the selection of p_i and q_i is a nontrivial problem. It is also desirable that the selection procedure be computationally simple in the sense that it may use add, subtract, and shift operations only. In general, this requires the use of an approximation in the selection procedure [2].

A selection procedure was obtained for the solution of a quadratic equation [2]. This was later extended to higher degree polynomials [4]. In this correspondence, we show that for functions expandable using the Riccati equation approach and for the function $\log_e x$, a simple selection procedure does not exist.

In Section II, we derive the expansions of functions into continued fractions. In Section III, we investigate the selection problem.

II. METHODS OF EXPANSION

Let the function to be expanded into a continued fraction be denoted by $f(a_0)$ where a_0 is a vector of arguments. We expand $f(a_i)$ (for $i = 0, 1, 2, \dots$) using the following bilinear transformation:

$$f(a_i) = \frac{p_{i+1}}{q_{i+1} + f(a_{i+1})}. \quad (2.1)$$

It is required that the vector of coefficients a_{i+1} be obtainable from a_i , p_{i+1} , q_{i+1} , a_{i-1} , p_i , and q_i by means of simple recursions. A recursion is said to be simple if it uses shift, addition, and subtraction operations only. Let us denote this system of recursions by

$$a_{i+1} = G(a_i, a_{i-1}, p_{i+1}, p_i, q_{i+1}, q_i).$$

Next we show that many functions fall in this category.

Manuscript received January 20, 1976; revised February 10, 1977. This work was supported in part by the National Science Foundation under Grant NSF DCR 73-07998.

The author was with the University of Illinois at Urbana-Champaign, Urbana, IL 61801. He is now with the Department of Computer Science, Duke University, Durham, NC 27706.

A. Solution of a Quadratic Equation [2]

Let $\mathbf{a}_i = (b_i, c_i)$, $f(\mathbf{a}_i) = c_i(b_i + x)$ and $x = c_0/(b_0 + x)$ then $f(\mathbf{a}_0)$ is a solution to the quadratic $x^2 + b_0x - c_0 = 0$. In [2], a system of simple recursions \mathbf{G} is derived, which may be written as

$$\begin{aligned} b_{i+1} &= q_{i+1}c_i - q_ic_{i-1} + b_{i-1} \\ c_{i+1} &= q_{i+1}(b_i - b_{i+1}) + c_{i-1}. \end{aligned}$$

In [4], this method has been extended to higher degree polynomials.

Another method of expansion for the solution of a quadratic equation $b_0x_0^2 + c_0x_0 - d_0 = 0$ is obtained by letting $\mathbf{a}_i = (b_i, c_i, d_i)$, $f(\mathbf{a}_i) = x_i$ where $b_ix_i^2 + c_ix_i - d_i = 0$. Applying the transformation (2.1), the system \mathbf{G} can be written as

$$\begin{aligned} b_{i+1} &= d_i/p_{i+1}^2 \\ c_{i+1} &= 2d_i \frac{q_{i+1}}{p_{i+1}^2} - \frac{c_i}{p_{i+1}} \\ d_{i+1} &= b_i + \frac{c_i q_{i+1}}{p_{i+1}} - d_i \left(\frac{q_{i+1}}{p_{i+1}} \right)^2. \end{aligned}$$

B. Expansion of Logarithm

Let $\mathbf{a}_i = (b_i, b_{i-1})$ and $f(\mathbf{a}_i) = \log_{b_{i-1}} b_i$. Applying the transformation (2.1), we have [5],

$$b_{i+1} = \frac{(b_{i-1})^{p_{i+1}}}{(b_i)^{q_{i+1}}}. \quad (2.2)$$

However, we note that this recursion is not simple. To solve this problem we can easily establish by induction that [6]

$$b_i = \left(\frac{(b_{-1})^{c_i}}{(b_0)^{d_i}} \right)^j, \quad (2.3)$$

where $j = 1$ if i is odd, $j = -1$ if i is even and the recursions for c_{i+1} and d_{i+1} are

$$\left. \begin{aligned} c_{-1} &= d_0 = 1, c_0 = d_{-1} = 0 \\ c_{i+1} &= p_{i+1}c_{i-1} + q_{i+1}c_i \\ d_{i+1} &= p_{i+1}d_{i-1} + q_{i+1}d_i. \end{aligned} \right\} \quad (2.4)$$

Comparing the recursion (1.1) and (2.4), we see that $c_i = P_i$ and $d_i = Q_i$ for all i . Therefore, if we let $\mathbf{a}_i = (P_i, Q_i)$, we have, $f(\mathbf{a}_0) = \log_{b_{-1}} b_0$.

C. The Riccati Equation [3], [7]

Consider the first-order differential equation:

$$y'_i + \sum_{j=-m}^n (a_i)_j y^j = 0. \quad (2.5)$$

We apply the bilinear transformation

$$y_i = p_{i+1}/(q_{i+1} + y_{i+1})$$

to (2.5) and require that y_{i+1} satisfy a similar differential equation, i.e.,

$$y'_{i+1} = \sum_{j=-m}^n (a_{i+1})_j y_{i+1}^j.$$

After some tedious algebra, it is easily shown that $m = 0$ and

$n = 2$. Now (2.5) is seen to be the well-known Riccati equation. Let

$$y'_i = j(a_i y_i^2 + b_i y_i + c_i), \quad (2.6)$$

where $j = 1$ if i is even, $j = -1$ if i is odd and the initial condition is, $y_i(0) = t_i$. Applying the bilinear transformation, we obtain the system of recursions [7]:

$$\left. \begin{aligned} a_{i+1} &= c_i/p_{i+1} \\ b_{i+1} &= b_i + 2c_i q_{i+1}/p_{i+1} \\ c_{i+1} &= a_i p_{i+1} + b_i q_{i+1} + c_i q_{i+1}^2/p_{i+1} \\ t_{i+1} &= p_{i+1}/t_i - q_{i+1}. \end{aligned} \right\} \quad (2.7)$$

Now if we let $\mathbf{a}_i = (a_i, b_i, c_i, t_i, x)$ and $f(\mathbf{a}_i) = y_i(x)$ then we have a method of expansion of $y_0(x) = f(\mathbf{a}_0)$. The system \mathbf{G} is given by the set of recursions (2.7). We note that the recursions for a_{i+1} , b_{i+1} , and c_{i+1} are simple since we have assumed that p_{i+1} and q_{i+1} are integral powers of 2. However, the recursion for t_{i+1} is not simple. This problem can be solved by letting $t_i = d_i/e_i$, $d_0 = t_0, e_0 = 1$ and

$$\begin{aligned} d_{i+1} &= k_{i+1}(p_{i+1}e_i - q_{i+1}d_i) \\ e_{i+1} &= k_{i+1}(d_i). \end{aligned} \quad (2.8)$$

We adjoin the recursions (2.8) to (2.7) after removing the recursion for t_{i+1} . Also, the vector \mathbf{a}_i is redefined so that $\mathbf{a}_i = (a_i, b_i, c_i, d_i, e_i, x)$. By choosing the initial coefficients a_0, b_0, c_0, d_0 , and e_0 appropriately, many different functions can be expanded using this approach, as shown in the following table [7].

a_0	b_0	c_0	t_0	$y_0(x)$
1	0	1	0	$\tan x$
-1	0	-1	∞	$\cot x$
-1	0	0	∞	$1/x$
-1	0	1	∞	$\coth x$
-1	0	1	0	$\tanh x$
0	+1	0	1	e^x
0	-1	0	1	e^{-x}

If we allow the coefficients a_i, b_i , and c_i to be functions of x then many more functions can be expanded [7].

We conclude this section by presenting an algorithm for the evaluation of a function using continued fractions. The problem of selection, which is hidden in the procedure "select" of the algorithm, will be discussed in the next section.

Algorithm A

Step 1 [Initialize]:

$$P_0 \leftarrow Q_{-1} \leftarrow 0; P_{-1} \leftarrow Q_0 \leftarrow 1; i \leftarrow 0;$$

Initialize the coefficient vector \mathbf{a}_0 depending on the function to be evaluated.

Step 2 [Selection]:

$$(p_{i+1}, q_{i+1}) \leftarrow \text{select}(\mathbf{a}_i, \text{function to be evaluated}).$$

Step 3 [Recursions]:

$$\mathbf{a}_{i+1} \leftarrow \mathbf{G}(\mathbf{a}_i, \mathbf{a}_{i-1}, p_i, p_{i+1}, q_i, q_{i+1});$$

$$P_{i+1} \leftarrow p_{i+1}P_{i-1} + q_{i+1}P_i;$$

$$Q_{i+1} \leftarrow p_{i+1}Q_{i-1} + q_{i+1}Q_i.$$

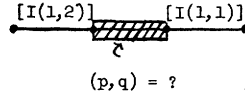


Fig. 1. The gap in selection.

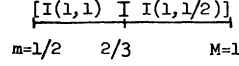


Fig. 2. A nonredundant number system.

Step 4 [Test]:

After a sufficient number of iterations GOTO Step 5; otherwise set $i \leftarrow i + 1$ and return to Step 2.

Step 5 [Evaluate]:

$$f(a_0) \simeq \frac{P_{i+1}}{Q_{i+1}};$$

End A.

III. THE SELECTION PROBLEM

Let the set of allowable values of partial numerators be denoted by S_p and the set of allowable values of partial denominators be denoted by S_q . We will assume that both S_p and S_q are finite subsets of positive reals. Let $p_{\min} = \min S_p$, $p_{\max} = \max S_p$, $q_{\min} = \min S_q$ and $q_{\max} = \max S_q$. Let the set of numbers representable as infinite continued fractions (ICF's) using the sets S_p and S_q be denoted by $R(S_p, S_q)$. Let

$$m = \frac{p_{\min}}{q_{\max} + \frac{p_{\max}}{q_{\min} + m}}$$

and let

$$M = \frac{p_{\max}}{q_{\min} + \frac{p_{\min}}{q_{\max} + M}}.$$

It is clear that

$$R(S_p, S_q) \subseteq [m, M].$$

We would like to impose some conditions on the sets S_p and S_q so that $R(S_p, S_q) = [m, M]$. As a result, any number in the interval $[m, M]$ can be represented as an ICF. Let $m(p, q) = p/(q + M)$, $M(p, q) = p/(q + m)$, $I(p, q) = [m(p, q), M(p, q)]$ and

$$I(S_p, S_q) = \bigcup_{\substack{p \in S_p \\ q \in S_q}} I(p, q).$$

Note that, $I(p, q)$ is a closed interval of the positive real numbers. It can be shown that the following theorem holds [6].

Theorem 1:

$$R(S_p, S_q) = [m, M] \quad \text{iff} \quad I(S_p, S_q) = [m, M].$$

Given the sets S_p and S_q , if the conditions of Theorem 1 are satisfied then we say that $R(S_p, S_q)$ is a number system (NS). Given an $f_0 \in [m, M]$ we can expand it into an ICF by letting

$$f_{i-1} = \frac{p_i}{q_i + f_i} \quad i = 1, 2, 3, \dots$$

The method of selection of the pair (p_i, q_i) is as follows.

Search for a pair (p_i, q_i) such that

$$p_i \in S_p, q_i \in S_q \text{ and } f_{i-1} \in I(p_i, q_i).$$

Note that this search will always succeed provided $R(S_p, S_q)$ is an NS. Furthermore, the definition of $I(p, q)$ guarantees that $f_i \in [m, M]$ therefore, the above procedure can be applied repetitively.

As an example, let $S_p = \{1\}$ and $S_q = \{1, 2\}$. In this case a simple computation reveals that the conditions of Theorem 1 are not satisfied and $R(S_p, S_q)$ is not an NS. The gap between selection intervals $I(1, 1)$ and $I(1, 2)$ is the reason for trouble as shown by Fig. 1.

As another example, let $S_p = \{1\}$ and $S_q = \{1, 1/2\}$. In this case there are no gaps as shown by Fig. 2.

Therefore, $R(S_p, S_q)$ forms an NS. In this case, the selection procedure can be specified as follows:

- If $f_{i-1} \in [1/2, 2/3)$, then $p_i = 1, q_i = 1$.
- If $f_{i-1} \in (2/3, 1]$, then $p_i = 1, q_i = 1/2$.
- If $f_{i-1} = 2/3$, then $p_i = 1$ and $q_i = 1/2$ or 1.

Note that two choices are possible for q_i if $f_{i-1} = 2/3$. Let an interval $I(p, q)$ be known as a selection interval. The reason for multiple choice is seen to be the nonnull intersection of adjacent selection intervals. As a result of this, some numbers in $[m, M]$ will have multiple ICF representations. Let us define an NS $R(S_p, S_q)$ to be nonredundant provided for any two distinct pairs (p, q) and (p', q') , $I(p, q) \cap I(p', q')$ is either null or is a singleton. In such a case it is easy to see that multiple choice of (p_i, q_i) results for only a finite number of points $f_{i-1} \in [m, M]$. We see that for $S_p = \{1\}$ and $S_q = \{1, 1/2\}$, $R(S_p, S_q)$ is a nonredundant NS. An example of a redundant NS is obtained by letting $S_p = \{1\}$ and $S_q = \{1, 1/2, 1/4\}$. In this case we note that [8],

$$I(1, 1/2) \cap I(1, 1) = [0.485, 0.72]$$

and

$$I(1, 1/2) \cap I(1, 1/4) = [0.553, 1.124].$$

Thus far, we have outlined a selection procedure when the number to be expanded is known explicitly. However, when using the algorithm of Section II, the number to be expanded at the i th step (i.e., $f(a_i)$) is known only implicitly via the coefficient vector \mathbf{a}_i . Therefore, we should specify the selection procedure in terms of \mathbf{a}_i . Recall that in terms of $f(a_i)$, the condition for selecting $(p_{i+1}, q_{i+1}) = (p, q)$ is that $f(a_i) \in I(p, q)$. This condition must, somehow, be translated in terms of \mathbf{a}_i . Even after such a transformation, it turns out that a prohibitive amount of computation is needed in the selection procedure. We may, however, reduce the computation by use of an approximation. By using a redundant NS, we hope that the error introduced in the selection due to the use of an approximation will be corrected by the redundancy of the NS.

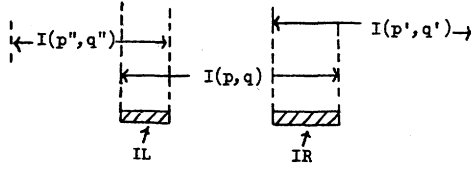


Fig. 3. Redundant number system.

A. Selection for the Quadratic [2], [8]

The (p, q) selection condition can be written as

$$m(p, q) \leq \frac{c_i}{b_i + x} \leq M(p, q)$$

or

$$(b_i + x)m(p, q) \leq c_i \leq (b_i + x)M(p, q). \quad (3.1)$$

Note that, $f(a_0) = x$ is the unknown to be expanded, therefore, we must use an approximation to x . Let us assume that three adjacent selection intervals $I(p, q)$, $I(p', q')$, and $I(p'', q'')$ are as shown in Fig. 3.

Thus, IL and IR are selection intersection intervals. Let us assume that we have an approximation \tilde{x} of x (and \tilde{x} is simple to compute from b_0 and c_0), and zl and zr are properly chosen constants such that $zl \in IL$ and $zr \in IR$. We may now use the following (p, q) selection rule:

$$(b_i + \tilde{x}) * zl \leq c_i \leq (b_i + \tilde{x}) * zr. \quad (3.2)$$

It is clear that the selection rule (3.2) may only be used provided the interval of c_i specified by (3.2) is contained in the interval specified by (3.1). In other words,

$$(b_i + x) * m(p, q) \leq (b_i + \tilde{x}) * zl$$

and

$$(b_i + \tilde{x}) * zr \leq (b_i + x) * M(p, q).$$

Thus, we have a restriction on the maximum error allowable in approximating x by \tilde{x} . In [2], [8], an approximation \tilde{x} satisfying these conditions was derived. Thus, we have an algorithm for the solution of a quadratic equation. This was later extended to higher degree polynomials.

Selection for the second method of the solution to a quadratic is even simpler. Since the (p, q) selection rule in terms of x_i is that $x_i \in I(p, q)$. An approximation \tilde{x}_i to x_i can easily be obtained from the coefficients b_i , c_i , and d_i .

B. Selection for Other Functions

In Section III-A we have shown that the selection problem can be solved for the roots of a quadratic and higher degree polynomials. This is the only class of problems for which the selection problem has been solved. For the remaining functions that we discussed in Section II, it is possible to show that no simple selection procedure exists. We will outline an intuitive proof of this contention; for rigorous proofs the reader may consult [6], [9], and [10].

Recall that in terms of $f(a_i)$, the (p, q) selection condition is that $f(a_i) \in I(p, q)$. Since $f(a_i)$ is an unknown, this must be translated into a (p, q) selection condition for a_i . Such a selection

condition will, clearly, require the computation of the inverse function f^{-1} . Since the computation of f^{-1} is generally as complex as the computation of f , we require that an approximation of f^{-1} be used in the selection procedure. Thus the whole process of evaluating f may be looked upon as an attempt to obtain a good approximation of f given a crude approximation of f^{-1} . It is hoped that the redundancy in number representation will allow us to make a correct choice of the coefficients in spite of this approximation.

Let us split the coefficient vector a_i into two vectors so that $a_i = (\alpha_i, \beta)$. Thus, the vector α_i consists of all the coefficients which vary with the index i and β consists of the invariant coefficients. As an example, in the case of the quadratic, $\alpha_i = (b_i, c_i)$ and β is null. As another example, for the Riccati approach, $\alpha_i = (a_i, b_i, c_i, d_i, e_i)$ and $\beta = (x)$. We say that the initial coefficient vector a_0 together with the system of recursions G determine the function to be evaluated and β is the vector of true arguments for which the function is to be evaluated. Note that β will play a role in the selection procedure. Since we have assumed that an approximation to f^{-1} is used in the selection procedure, we can find two values of β , namely, β_1 and β_2 , such that $\beta_1 \neq \beta_2$ but the corresponding approximation of f^{-1} yields the same value. Note that since $(\alpha_0)_1 = (\alpha_0)_2$, we have that $(p_1, q_1)_1 = (p_1, q_1)_2$. With this condition we can prove by induction that $(\alpha_i)_1 = (\alpha_i)_2$ and $(p_i, q_i)_1 = (p_i, q_i)_2$ for all i . Therefore, $f(\alpha_0, \beta_1) = f(\alpha_0, \beta_2)$. Thus f is not able to resolve β values if the approximation to f^{-1} is not able to resolve the same β values. It is therefore clear that for our procedure to work, we must require that the β vector be null. Indeed, in the case of the solution to polynomial equations β vector is null. In the Riccati Approach, β vector is always nonnull. In the unmodified expansion of $\log_{b-1} b_0$, $\alpha_i = (b_i, b_{i-1})$ and β is null. But since the system G was not simple, we applied a transformation to obtain $\alpha_i = (P_i, Q_i)$ and $\beta = (b_0, b_{-1})$. This makes the selection problem unsolvable.

IV. CONCLUSION AND FURTHER REMARKS

Recently, there has been some interest in the use of continued fractions for digital hardware calculations. We require that the coefficients of the continued fractions be integral powers of 2. As a result well-known continued fraction expansions of functions cannot be used. We have presented methods of expansion of a large number of functions into continued fractions.

Selection of coefficients of the continued fractions is, however, a difficult problem. We have shown that the selection problem can be solved for the solution of a quadratic and higher degree polynomial equations. However, this is the only class of problems for which the selection problem has been solved. We have shown that for most of the remaining functions discussed in this correspondence no simple selection procedure can be found [10].

ACKNOWLEDGMENT

The author wishes to thank Prof. J. E. Robertson for introducing him to the subject and for continued support and encouragement.

REFERENCES

- [1] H. S. Wall, *Analytic Theory of Continued Fractions*. Princeton, NJ: Van Nostrand, 1950.
 - [2] J. E. Robertson and K. S. Trivedi, "The status of investigations into computer hardware design based on the use of continued fractions," *IEEE Trans. Comput.*, vol. C-22, pp. 555-560, June 1973.
 - [3] P. Wynn, "On some recent developments in the theory and application of continued fractions," *J. SIAM Numer. Analysis*, vol. 1, pp. 177-197, 1964.
 - [4] A. Bracha, "A method for solving polynomial equations by continued fractions," *IEEE Trans. Comput.*, vol. C-23, p. 1093, Oct. 1974.
 - [5] L. A. Lyusternik et al., *Handbook for Computing Elementary Functions*. New York: Permagon Press, 1965.
 - [6] K. S. Trivedi, "On a negative result regarding the use of continued fractions for digital computer arithmetic," Univ. of Illinois, Dep. of Comput. Sci. Report UIUCDCS-R-75-693, Jan. 1975.
 - [7] —, "The use of Riccati equation in digital computer arithmetic," University of Illinois, Dep. Comput. Sci. Rep. UIUCDCS-R-74-674, Aug. 1974.
 - [8] —, "An algorithm for the solution of a quadratic equation using continued fractions," M.S. thesis, Univ. of Illinois, Urbana, June 1972; also Department of Computer Science Rep. 525.
 - [9] —, "Further negative results regarding the use of continued fractions for digital computer arithmetic," Univ. of Illinois, Dep. of Comput. Sci. Rep. UIUCDCS-R-75-721, May 1975.
 - [10] —, "On the use of continued fractions for digital computer arithmetic," in *Proc. 3rd IEEE Symp. on Computer Arithmetic*, Dallas, TX, Nov. 1975.
-