Radix-16 Evaluation of Certain Elementary Functions

MILOŠ D. ERCEGOVAC

Abstract—This paper describes a family of algorithms for evaluation of a class of elementary functions including division, logarithms, and exponentials. The main objective is to demonstrate the feasibility of higher radix implementations, in particular, radix 16, and to compare performance with radix 2. The emphasis is not on optimality of a single algorithm, but rather on the optimality of a class of algorithms. An attempt to implement a much wider class of functions than is presently done in arithmetic units is encouraged by the current level of digital technology and the existence of suitable algorithms. Besides the definitions of the algorithms, which are based on continued products and continued sums, details related to implementation are discussed.

Index Terms—Continued products, continued sums, division, exponential, logarithm, pipelining, radix 16, redundant number, symmetric digit set.

I. INTRODUCTION

The use of continued products and sums has long been recognized as a practical way to reduce the evaluation of certain elementary functions to a series of additions, shift operations, and perhaps access to precomputed constants [1]-[10]. Recently, DeLugish [1] defined efficient algorithms for the evaluation of a wide class of elementary functions based upon continued products and sums, and the use of radix-2 redundant digit set \{1, 0, 1\} (I = -1). The algorithms generate the value of the function on a digit-by-digit basis. Redundancy [13] simplifies the digit selection procedure and accelerates the computation by increasing the probability of a zero digit value. DeLugish has shown that for a given function, including division, multiplication, square root, logarithm, exponential, trigonometric, and inverses trigonometric functions, operation times are from one to three multiplication cycle times.

In this paper, we are concerned with speeding up some of the DeLugish algorithms using the radix-16 digit-by-digit evaluation, i.e., 4 bits/step. Only the algorithms for the fractional parts of the radix-2 floating-point numbers are considered. The arguments are assumed to be of \(m = \frac{M}{4}\), radix-16 digits precision. By the multiplicative normalization of a given argument \(X_0 \in [\frac{1}{2}, 1)\), we mean a sequence of transformations such that \(X_0 \cdot \frac{m}{M_0} M_1 \to 1\) where the multipliers are of the form \(M_k = 1 + S_k \cdot 16^{-k}\) for \(0 < k < m\). Similarly, in the additive normalization, \(X_0 - \sum_{i=0}^{m} A_i = 0\) with \(A_k = S_k \cdot 16^{-k}\). In both cases, \(S_k \in \{10, \cdots, 10\}\). As defined by Robertson [12], the set of values for \(S_k\) with the redundancy ratio of \(\frac{3}{2}\) allows an efficient multiple formation as well as low precision selection rules.

II. MULTIPLICATIVE NORMALIZATION

The multiplicative normalization is performed recursively as

\[ X_{k+1} = X_k (1 + S_k \cdot 16^{-k}), \quad 0 \leq k \leq m, \quad (2.1) \]

normalizing \(X_0\) to 1 in \(m + 1\) steps, by appropriately choosing the values of \(S_k\). \(S_k\) can be determined on the basis of \(X_k\), but to keep selection dependent on the same register positions, the scaled remainder is defined as

\[ R_k = 16^{k-1} (X_k - 1), \quad 0 < k < m. \quad (2.2) \]

From (2.1) and (2.2), the scaled remainder recursion follows:

\[ R_{k+1} = 16R_k + S_k + S_k R_k \cdot 16^{-k+1}, \quad 0 < k < m. \quad (2.3) \]

For the chosen symmetric redundant digit set \(\{S_k\}\) with max \(\{S_k\} = 10\), \(R_{k+1}\) will remain in the same bounds as \(R_k\) if \(-\frac{3}{2} < S_k < \frac{3}{2}\). Correspondingly, \(1 - (\frac{3}{2}) \cdot 16^{-k} < X_k < 1 + (\frac{3}{2}) \cdot 16^{-k}\). The first step in the derivation of the selection rules, as described in detail in [17], is the calculation of the intervals in the range of \(R_k\) for all \(k\) and for each value of \(S_k\) such that \(R_{k+1}\) satisfies the bounds. To preserve continuity of the \(R_k\) range, the allowed values for \(S_k\) must correspond to the overlapping intervals in the \(R_k\) range. These overlaps are caused by the redundancy in the digit set \(\{S_k\}\). Since it is possible to choose for the interval boundaries the numbers that are simple in the binary sense, the low precision operations will suffice in the implementation of the selection rules. Since the correspondence between values of \(S_k\) and the intervals of \(R_k\) reduces, for \(2 < k < m\), to

\[ (-2S_k - 1) \leq 32R_k \leq (-2S_k + 1), \quad (2.4) \]

a simple selection rule will be to form \(S_k\) as the scaled remainder rounded to the most significant digit with the opposite sign. Then, it is convenient to perform selection in the first three “irregular” steps using modified rounding instead of using a table lookup or a direct combinational approach.

The following definitions are used in the algorithms:

1) Two’s complement representation of scaled remainders:

\[ R_k = -r_0 + \sum_{i=1}^{4m} r_i 2^{-i}, \quad r_i \in \{0, 1\} \text{ for all } i. \quad (2.5) \]

2) Truncated scaled remainder:

\[ \hat{R}_k = -r_0 + \sum_{i=1}^{k} r_i 2^{-i}. \quad (2.6) \]

3) Nonsign part of \(\hat{R}_k\):
\[ T_k = \begin{cases} \sum_{i=1}^{k} r_i 2^{-i}, & \text{if } r_0 = 0 \\ \sum_{i=1}^{k} \bar{r}_i 2^{-i}, & \text{if } r_0 = 1. \end{cases} \]  

(2.7)

4) Step-dependent rounding constant:

\[ U_k = \sum_{i=1}^{k} u_i 2^{-i}, \quad u_i \in \{0, 1\} \text{ and } u_i = f_i(\bar{R}_k). \]  

(2.8)

5) \( S_k \) is in sign and magnitude representation.

6) \([Z]\) denotes the integer part of \( Z \).

7) \([Z]\) denotes the smallest integer not smaller than \( Z \).

(2.9)

(2.10)

Algorithm \( N \) (Multiplicative Normalization)

\textbf{Step N1:} (Initialize) \( k \leftarrow 0, \)

\[ S_0 \leftarrow 1 \text{ if } \frac{1}{2} < X_0 < \frac{5}{2}, \]

\[ S_0 \leftarrow 0 \text{ if } \frac{5}{2} < X_0 < 1, \]

\[ R_1 \leftarrow X_0 (1 + S_0) - 1. \]

\textbf{Step N2:} (Loop) \( \) for \( 0 < k \leq m: \)

\[ k \leftarrow k + 1, \]

\[ |S_k| \leftarrow [(T_k + U_k)16]; \text{ sign } (S_k) \leftarrow \text{ sign } (R_k), \]

if \( k < k_1 \), then

\[ R_{k+1} \leftarrow 16R_k + S_k + S_kR_k \cdot 16^{-k+1}; \]

otherwise \( R_{k+1} \leftarrow 16R_k + S_k \)

where

\[ U_k = \sum_{i=1}^{k} u_i 2^{-i}, \]

\[ u_1 = u_2 = 0, \]

\[ u_3 = K_1r_0 r_2, \]

\[ u_4 = K_1 r_0 r_4 (\bar{r}_2 + \bar{r}_3), \]

\[ u_5 = K_1 (r_0 + \bar{r}_2 \bar{r}_4) + K_2 [r_0 + \bar{r}_2 (\bar{r}_3 + r_6)] + K, \]

\[ u_6 = K_1 r_3 r_4 + K_2 r_0 (r_1 + r_2 r_3), \]

and \( K_1, K_2, \) and \( K \) denote steps \( k = 1, k = 2, \) and \( k > 2 \) respectively.

b) \( k_1 = \lceil (m + 3)/2 \rceil. \)

This simplification is the consequence of the decreasing effect of the term \( S_k R_k 16^{-k+1} \) on \( R_{k+1} \). It is a useful feature of this normalization procedure that at every step, an increasing number of digits \( S_k \) can be obtained from \( R_k \). This reveals the possibility of a variable radix approach to achieve higher speed. At the step \( k = k_1 \), all remaining digits \( S_k, S_{k+1}, \ldots, S_m \) are available after replacing the leading digits of \( R_k \), with the digits from the set \( \{8, \cdots, 8\} \), in such a way that the value of the truncated remainder remains unchanged. This also reduces the error bound of the normalized argument \([14]\) to

\[ |e_{m+1}| = 1 - X_{m+1} | \leq (\frac{8}{15}) \cdot 16^{-m}. \]  

(2.12)

III. ADDITIVE NORMALIZATION

The additive normalization is a right-directed recoding where one replaces the nonredundant digit set \( \{0, \cdots, 15\} \) with a redundant one \( \{10, \cdots, 15\} \). The procedure is simple and exact. Rounding, as a selection rule, applies to all steps. Since \( X_{k+1} = X_k - S_k 16^{-k}, \) \( 0 \leq k \leq m \) \n
(3.1)

with the scaled remainder defined as \( R_k = 16^{-k-1} X_k \), the basic scaled remainder recursion is

\[ R_{k+1} = 16R_k - S_k, \quad 0 < k \leq m \]  

(3.2)

and \( |R_k| \leq \frac{1}{16}. \)

Algorithm A (Additive Normalization)

\textbf{Step A1:} (Initialize) \( k \leftarrow 0, \)

\[ S_0 \leftarrow 1, \]

\[ R_1 \leftarrow X_0 - S_0, \]

\textbf{Step A2:} (Loop) \( \) for \( 0 < k \leq m: \)

\[ k \leftarrow k + 1, \]

\[ |S_k| \leftarrow [(T_k + U_k)16]; \text{ sign } (S_k) \leftarrow \text{ sign } (R_k), \]

\[ R_{k+1} \leftarrow 16R_k - S_k \]

where \( U_k = \frac{1}{2}. \) This choice of the rounding constant restricts \( S_k \) to \( \{8, \cdots, 8\} \), which is sufficient for recoding.

IV. DIVISION

Consider

\[ Q = \frac{Y_0}{X_0} = \prod_{i=0}^{m} M_i \]

(4.1)

where \( M_k = 1 + S_k \cdot 16^{-k}, \) \( 0 \leq k \leq m. \) If the digits \( S_k \) are determined using the multiplicative normalization (Algorithm \( N \)), then

\[ X_0 \prod_{i=0}^{m} M_i = X_{m+1} = 1 + e_{m+1} \]

(4.2)

where \( |e_{m+1}| \leq \left(\frac{8}{15}\right) \cdot 16^{-m}, \) given in (2.12). Consequently,

\[ Y_0 \prod_{i=0}^{m} M_i = Q_{m+1} = Q \]

(4.3)

with the relative error \( |\delta| = |e_{m+1}|. \) The partial results are defined recursively as

\[ Q_{k+1} = Q_k (1 + S_k 16^{-k}), \quad 0 \leq k \leq m \]

(4.4)

\[ Q_0 = Y_0. \]

An example of the division algorithm is given in Fig. 1. The multiplicative normalization scheme is shown in Fig. 2, and the result evaluation scheme is shown in Fig. 3.

Algorithm D (Division)

\textbf{Step D1:} \( k \leftarrow 0; \)  

\textbf{Normalization:} \( \)  

\textbf{Step D2:} \( \) for \( k \leq m: \)

\[ \begin{array}{l}
\text{Algorithm } N; \\
\text{Evaluation:}
\end{array} \]

\[ Q_0 = Y_0; \]

\[ Q_{k+1} = Q_k + Q_k S_k 16^{-k}. \]

(4.5)

Note in Fig. 1 that the six leading digits of \( R_7 \) represent, after recoding, the remaining digits \( S_7, \cdots, S_{12}: 0.597\text{BEA}9 \cdots = 0.6758\text{F}1\cdots.\)
Fig. 1. Example of the division algorithm.

Fig. 2. Multiplicative normalization scheme.

Fig. 3. Result evaluation scheme.

Fig. 4. Result evaluation scheme with ROM.

\[
\ln X_0 = \ln \left( X_0 \prod_{i=0}^{m} M_i \right) - \ln \left( \prod_{i=0}^{m} M_i \right) 
\]

(5.2)

with \( M_k = 1 + S_k \cdot 16^{-k} \), \( 0 \leq k \leq m \). According to the Algorithm \( N \), \( \left| X_0 \prod_{i=0}^{m} M_i \right| \leq \max \{ 16, 1 \} \cdot 16^{-m} \). Then

\[
\ln X_0 \approx - \sum_{i=0}^{m} \ln (1 + S_i \cdot 16^{-i}).
\]

(5.3)

Since the logarithmic constants \( \ln (1 + S_k \cdot 16^{-k}) \) are stored with \( m \)-digits precision in a fast Read-Only memory (ROM), the result will contain an accumulated roundoff error \( \varepsilon_p \leq \left( \frac{1}{2} m \right) \cdot 16^{-m} \), inversely proportional to the radix of implementation. Due to the finite precision, the actual requirement for the ROM capacity is decreased by approximately one-half, as follows from the power series expansion for the logarithm:

\[
\ln (1 + S_k \cdot 16^{-k}) \approx S_k \cdot 16^{-k}
\]

(5.4)

for \( k \gg k_1 \) where \( k_1 = \lceil (2 \log_2 10 - 1 + 4m)/8 \rceil \).

See the multiplicative normalization scheme in Fig. 2, and the result evaluation scheme with ROM in Fig. 4.

Algorithm L (Logarithm)

(5.5)

Step L1: \( k \leftarrow 0 \); 
Step L2: for \( k \leq m \):

Normalization: [ Result Evaluation: ]

\[
L_0 = 0; \quad \text{if } k < k_1, \text{ then } L_{k+1} \leftarrow L_k - \ln(1 + S_k \cdot 16^{-k}); \text{ otherwise } L_{k+1} \leftarrow L_k - S_k 16^{-k}.
\]

VI. EXPONENTIAL

An initial manipulation \([3, 1]\) of the argument reduces the evaluation of \( e^X \) to the evaluation of \( e^{z_0} \) with \( |X_0| < \ln 2 \).

Let \( X \log_2 e = I + F \) where \( I \) and \( F \) denote the integer and fractional parts, respectively. Since \( X = X \log_2 e \ln 2 \),

\[
e^X = e^{I \log_2 e} \cdot e^F = 2^I \cdot e^F.
\]

(6.1)
The factor 2^l is easily incorporated into the exponent part of the result. To simplify the initial step, F is restricted to F < 0 and I is correspondingly adjusted so that X_0 ∈ (-ln 2, 0). Consider

\[ e^{X} = 2^l e^{F \ln 2} = 2^l e^{X_0}. \]  

(6.1)

The factor \(2^l\) is easily incorporated into the exponent part of the result. To simplify the initial step, \(F\) is restricted to \(F < 0\) and \(I\) is correspondingly adjusted so that \(X_0 \in (-\ln 2, 0)\). Consider

\[ X_0 = X_0 - \ln \left( \prod_{i=1}^{m} M_i \right) + \ln \left( \prod_{i=0}^{m} M_i \right) \]  

(6.2)

where \(M_k = 1 + S_k \cdot 16^{-k}, 0 \leq k \leq m\). If \(X_0 - \ln (\prod_{i=0}^{m} M_i) \approx 0\), then \(e^{X_0} \approx \prod_{i=0}^{m} M_i\). The corresponding result evaluation recursion is

\[ Q_0 = 1 \]

\[ Q_{k+1} = Q_k (1 + S_k \cdot 16^{-k}), \quad 0 \leq k \leq m. \]  

(6.3)

The constants \(S_k\) are determined by the additive normalization, using the scaled remainder recursion:

\[ R_{k+1} = 16R_k - 16^k \ln (1 + S_k \cdot 16^{-k}), \quad 0 \leq k \leq m. \]  

(6.4)

The set of stored constants is the same as the one for the logarithmic algorithm. For \(k \geq k_1\), (5.4), the normalization is defined by the Algorithm A, Step A2. As derived in [17], the rounding with \(U_k = \frac{1}{32}\) is the selection procedure in all steps, except in the initial, which is performed according to the following table.

<table>
<thead>
<tr>
<th>(X_0)</th>
<th>(M_0)</th>
<th>(\ln M_0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>([-\frac{1}{8}, 0])</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>([-\frac{1}{8}, -\frac{1}{2}))</td>
<td>(e^{-1/8})</td>
<td>(-\frac{1}{2})</td>
</tr>
<tr>
<td>((-\ln 2, -\frac{3}{8}))</td>
<td>(e^{-17/32})</td>
<td>(-\frac{17}{32})</td>
</tr>
</tbody>
</table>

The constants, necessary to be precomputed, are stored in the ROM. A summary of the procedure for calculation of \(e^X\) follows.

**Preparatory Operations P**

(6.5)

**Step 1:** \(N \leftarrow X \log_2 e\).

**Step 2:** If \(X > 0\), then \(I \leftarrow [N] + 1\); otherwise \(I \leftarrow [N]\).

**Step 3:** \(X_0 \leftarrow (N - I) \ln 2\).

The additive normalization scheme with ROM is shown in Fig. 5, and the result evaluation scheme is shown in Fig. 2.

**Algorithm E (Exponential)**

(6.6)

**Step 1:** \(k \leftarrow 0\); 

**Step 2:** for \(k \leq m\):

- if \(k < k_1\), then \(S_k \leftarrow ([T_k + U_k]16)\), \(|S_k| \leftarrow ([T_k + U_k]16)\), \(S_k \leftarrow 16R_k - 16^k \ln (1 + S_k 16^{-k})\), otherwise Step A2;

- \(Q_k \leftarrow \ln M_0\);

- \(Q_{k+1} \leftarrow Q_k + \ln (1 + S_k 16^{-k})\);

where \(k_1\) is defined in (5.4) and \(U_k = \frac{1}{32}\).

**VII. Implementation**

This investigation of the use of radix 16 in the implementation of the described algorithms has been motivated by the speed improvement over the radix-2 approach and by possible tradeoffs in the hardware requirements. A brief discussion of the radix-16 implementation and some basic comparisons with the radix-2 approach [1] follow, with more details given in [17]. The general configuration may contain two arithmetic units, as the fastest solution, or the normalization and the result evaluation can be performed in an overlapped mode, using one “pipelined” arithmetic unit. In either case, the main parts of the arithmetic unit are as follows. The three-input adder with the multiple formation networks is estimated to be twice as complex in the radix-16 case as the corresponding part in the radix-2 case. Speedwise, we estimate that the add times satisfy \(t_{a16} < 1.2t_{a2}\). The fast variable shifting network of the “barrel switch”-type [16], on the other hand, requires more than 30 percent more hardware in the radix-2 case. Speedwise, \(t_{sh} > 1.3t_{sh_{16}}\), where \(t_{sh}\) denotes the shifting delay. The selection procedure in the radix-2 case requires a simple 4-bit comparison. The radix-16 approach requires 7 bits of precision for selection, implementation of the 5 \(u_j = f(R_k)\) equations (2.11), costing less than 40 literals, as well as the additional inputs to the 7 significant positions of the adder. In both cases, the selection requirements are negligible in comparison with the rest of the unit. The control is also simple in both cases.

From the above considerations, the hardware requirement ratio for the radix 2 and the radix 16 is approximately 2:3. The basic cycle can be taken to be the same, since the add time is dominant over control, selection, and shifting time.

---

1 Two multiplications in the \(X_0\) calculation could be replaced by one division of \(X\) by \(\ln 2\) so that \(X = l/\ln 2 + X_0\) if the division scheme gives the remainder [2], [11].
The ROM capacity requirement ratio is about 1:3 in favor of the binary case.

Let the performance of an implementation in the radix $r$ be $P_r = \log_2 r/T_r$, where $T_r$ is the total delay necessary to evaluate $\log_2 r$ bits of the result, as defined in [15]. $T_r$ is equivalent to the basic cycle. In the radix-2 case with $S_k\in\{1,0,1\}$, the probability of $S_k = 0$ ($p_0 = \frac{1}{3}$) is utilized by providing an adder bypass and reducing the number of full basic cycles to $M/3$ on the average, where $M$ is the number of bits. Then it can be taken that the radix-16 basic cycle is $T_{16} \approx 3T_2$, since the number of basic cycles in the radix-16 case is always the same, the probability of $S_k = 0$ being too low. Then the ratio of performances is $P_{16}/P_2 \approx \frac{3}{1}$ on the average. If the efficiency of the implementation is defined as the ratio between performance and cost per bit, then with all previous assumptions, $E_{16}/E_2 \approx 1$ without considering ROM requirements. If the ROM capacity requirement is taken into account, then the radix-2 approach will offer more efficient design, but the radix-16 case will maintain better performance with shorter execution time. The selection procedure for radix 16 has been shown to be sufficiently simple.

One way to overlap the normalization and the result evaluation using pipelining techniques is shown in Fig. 6. The timing chart (b) corresponds to the division algorithm. The superscripts ' and " denote the right and the left parts of all networks, registers, and corresponding results. Two periods now define the basic cycle, somewhat longer so that the pipelined mode is 15-25 percent slower than the two separate arithmetic unit mode. The adder with multiple formation networks is split into two equal parts, $AS''$ and $AS'$, by breaking the carry path and inserting a one-bit carry register $C$. Outputs from the left half $SN''$ of the shifting network are to be saved in a latch $L$. The selection is carried out in the block $S$ on the basis of adder outputs and returns the value of $S_k$. The initial operands are in register $B$ for normalization, and in register $A$ for result evaluation. Each register contains two separately controlled halves: $B''$, $B'$ and $A''$, $A'$. The outputs from $AS''$ and $AS'$ are connected, under a separate control, to the inputs of $A''$ and $A'$ registers, respectively. A more detailed description can be found in [17].

VIII. CONCLUSIONS

The use of a higher radix in implementations makes faster execution of the algorithms possible, but feasibility of the corresponding selection procedure requires some consideration. For the radix-16 case, we have shown that the selection rules are very simple for all except the initial three steps, in which case they are of an acceptable complexity. Furthermore, in the exponential algorithm, only the initial step requires a specialized procedure. With the current trend in lower hardware cost, the corresponding realization requirements can be easily justified. It is believed that similar selection procedures could be applied for square-root, trigonometric, and inverse trigonometric function evaluation in radix 16. Multiplication has been made compatible [17].

Two additional 4-bit latches, not shown here, are needed to hold bits shifted left from $AS'$ and $SN''$. 
ACKNOWLEDGMENT

The author wishes to thank Prof. J. E. Robertson of the University of Illinois, Urbana, for highly valued guidance and discussions. Thanks are also due to the referees for helpful comments.

REFERENCES


Miloš D. Ercegovac was born in Belgrade, Yugoslavia, on June 28, 1942. He received the diploma in electrical engineering from the University of Belgrade, Belgrade, Yugoslavia, in 1965, and the M.S. degree in computer science from the University of Illinois, Urbana, in 1972.

From February 1966 to June 1970 he was with the Institute for Automation and Telecommunications "Mihailo Pupin," Belgrade, Yugoslavia, where he participated in the design and development of digital computers. Since July 1970 he has been a Research Assistant in the Department of Computer Science, University of Illinois, Urbana. His research interests include computer arithmetic, logical design, and computing systems organization. He is currently a Ph.D. student in computer science at the University of Illinois.