BKM: a new hardware algorithm for complex elementary functions
Jean-Michel Muller, Jean-Claude Bajard, Sylvanus Kla

To cite this version:
Jean-Michel Muller, Jean-Claude Bajard, Sylvanus Kla. BKM: a new hardware algorithm for complex elementary functions. IEEE Transactions on Computers, 1994, 43 (8), pp.955-963. 10.1109/12.295857. ensl-00086894

HAL Id: ensl-00086894
https://ens-lyon.hal.science/ensl-00086894
Submitted on 20 Jul 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
BKM: A New Hardware Algorithm for Complex Elementary Functions

Jean-Claude Bajard, Sylvanlus Kla, and Jean-Michel Muller, Member, IEEE

Abstract—A new algorithm for computing the complex logarithm and exponential functions is proposed. This algorithm is based on shift-and-add elementary steps, and it generalizes some algorithms by Briggs and De Lugish, as well as the Cordic algorithm. It can easily be used to compute the classical real elementary functions (sin, cos, arctan, ln, exp). This algorithm is more suitable for computations in a redundant number system than the Cordic algorithm, since there is no scaling factor when computing trigonometric functions.

Index Terms—Computer arithmetic, elementary functions, cordic, redundant number systems.

I. INTRODUCTION

The point at stake in this paper is the search for algorithms that rapidly compute elementary functions. Many methods have been proposed in the literature: approximation by polynomials or rational functions [3], [5], [6], [10], use of Newton’s method, approximation by continued products, continued fractions, E-Method [8], [9] and shift-and-add methods.

The algorithms presented in this paper belong to the class of shift-and-add methods. These methods are based on simple elementary steps: additions, and shifts (i.e., multiplications by a power of the radix of the number system used), and they go back to the 17th century: Briggs, a contemporary of Neper, invented an algorithm that made it possible to build the first tables of logarithms. For instance, to compute the logarithm of $x$ in radix-2 arithmetic, numerous methods (including that of Briggs, adapted to this radix) [4], [12], [13], [15] broadly consist of finding a sequence $d_k = -1,0,1$, such that $x = \prod_{k=1}^{n} (1 + d_k 2^{-k}) \approx 1$. Then $\ln(x) \approx -\sum_{k=1}^{n} \ln(1 + d_k 2^{-k})$. Another method belonging to the shift-and-add class is the CORDIC algorithm, introduced in 1959 by J. Volder [17] and then generalized by J. Walther [18]. CORDIC consists of the following iteration:

$$
\begin{align*}
    x_{n+1} &= x_n - m d_n y_n 2^{-\sigma(n)} \\
    y_{n+1} &= y_n + d_n x_n 2^{-\sigma(n)} \\
    z_{n+1} &= z_n - d_n \sigma(n),
\end{align*}
$$

(1)

where $m$ equals 0, 1 or $-1$, and $d_n$ is equal to 1 or $-1$. The results and the values of $d_n$ are stored in Table I.

CORDIC allows computation of many functions [18]. For instance, $e^x$ is obtained as $\cosh x + \sinh x$, while $\ln x$ is obtained as $2 \tanh^{-1} \left| \frac{1-x}{1+x} \right|$. CORDIC has been implemented in many pocket calculators and floating-point coprocessors. Its major drawback arises when performing the iterations using a redundant (e.g., carry save or signed-digit) number system. Such systems are advantageous for quickly-performed arithmetic, since they make it possible to perform carry-free additions [2]. In these systems, $d_n$ is difficult to evaluate. For instance, assume that we are in the iteration mode of CORDIC (see Table I), and that numbers are represented by $p$ digits, in radix 2 with digits $-1,0$ or 1. $d_n$ is equal to the sign of the most significant nonzero digit of $z_n$: to find its value, we must examine some number of digits which may be close to $p$. Thus, the advantage of the redundant representation (a constant time elementary step) would be lost. An alternative is to accept $d_n = 0$: this makes it possible to examine a few digits of $z_n$ only. Unfortunately, with such a method the scale factors $K$ and $K'$ are not constants. $K$ is equal to $\prod_{n=0}^{\infty} \sqrt{1 + d_n 2^{-2n}}$, it is a constant if the $d_n$’s are all equal to $\pm 1$, but it is no longer a constant if the $d_n$’s are allowed to be zero. Many solutions have been suggested to overcome this problem. Broadly speaking, they lead to a repetition of iterations in time [16], [1], or in space [7]. In order to avoid this wasteful repetition, we need to work out a new algorithm.

In the following, assume that we are using a radix-2 classical or signed-digit number system. Extension to binary carry-save representation is straightforward. Let us consider the basic step of CORDIC in trigonometric mode (i.e., iteration (1) with $m = 1$). If we define the complex number $L_n$ as $L_n = x_n + i y_n$, we obtain $L_{n+1} = L_n (1 + i d_n 2^{-n})$, this relation is close to the basic step of Briggs’ algorithm. This remark brings us to a generalization of that algorithm: we could perform multiplications by terms of the form $1 + d_n 2^{-n}$, where the $d_n$’s are complex numbers, chosen such that a multiplication by $d_n$ can be reduced to a few additions. In this paper, we study the following iteration, called BKM:

$$
\begin{align*}
    L_{n+1} &= L_n (1 + d_n 2^{-n}) \\
    E_{n+1} &= E_n - \ln(1 + d_n 2^{-n})
\end{align*}
$$

(2)

with $d_n \in \{-1,0,1,-i,i,1+i,-1+i,-1-i\}$. We define $\ln z$ as the number $t$ such that $e^t = z$, and whose imaginary part is between $-\pi$ and $\pi$.

- If we are able to find a sequence $d_n$ such that $L_n$ goes to 1, then we will obtain $E_n \rightarrow E_0 + \ln(L_1)$. We call this iteration mode the L-mode of the BKM algorithm.

Manuscript received September 17, 1993; revised February 23, 1994.

J.-C. Bajard is with Lab. LIM, Université de Provence, Marseille, 3 Place Victor Hugo, 13331 Marseille Cedex 03, France.

J.-M. Muller is with CNRS, Lab. LIP, Ecole Normale Supérieure de Lyon, 46 Allée d’Italie, 69364 Lyon Cedex 07, France; e-mail: jmmuller@lip.ens-lyon.fr.

IEEE Log Number 9203096.
rithms for computing the complex logarithm and exponential functions.

We shall deduce convenient algorithms for computing the complex logarithm and exponential functions.

In this section, we give an algorithm which computes the complex logarithm and exponential functions.

II. COMPUTATION OF THE EXPONENTIAL FUNCTION (E-MODE)

As pointed out at the end of the previous section, for computing $e^{z}$ using BKM, one needs to find a sequence $d_n$ such that $E_n = L_1 e^y$. We call this iteration the E-mode of BKM.

So, in the next sections, we will focus on the problem of finding sequences $d_n$ such that $E_n$ goes to 1 or such that $E_n$ goes to 0. From this study, we will deduce convenient algorith-
m for computing the complex logarithm and exponential functions.

If we are able to find a sequence $d_n$ such that $E_n$ goes to 0, then we will obtain $L_n = L_1 e^{y}$. We call this iteration the E-mode of BKM.

So, in the next sections, we will focus on the problem of finding sequences $d_n$ such that $E_n$ goes to 1 or such that $E_n$ goes to 0. From this study, we will deduce convenient algorithms for computing the complex logarithm and exponential functions.

II. COMPUTATION OF THE EXPONENTIAL FUNCTION (E-MODE)

As pointed out at the end of the previous section, for computing $e^{z}$ using BKM, one needs to find a sequence $d_n$, $d_n = -1, 0, 1, -i, i, -1, 1, -i, 1, i, -1, -i, 1, i, -1, 1, i, -1, -i, 1, i, -1, -i$ such that $\lim_{n \to \infty} E_n = 0$. Let us define $d_k$ and $d_k'$ as the real and imaginary parts of $d_n$ (they belong to $\{-1, 0, 1\}$) and $E_n$ and $E_{n+1}$ as the real and imaginary parts of $E_n$. We easily find:

$$E_{n+1} = E_n - \frac{1}{2} \ln \left[ 1 + d_n^2 e^{-2^{n+1}} \right] + \left( d_n^2 + d_n^2 \right) 2^{-2n}$$

\[E_{n+1} = E_n - d_n \arctan \left( \frac{2^{-n}}{1 + d_n^2} \right)\]

(3)

In this section, we give an algorithm which computes the sequence $d_n$ for any $E_1$ belonging to a rectangular set $R_1 = [-s_{n+1}, r_{n+1}] + i[-r_{n+1}, r_{n+1}]$. The proof of our algorithm is based on the construction of a sequence $R_n = [-s_n, r_n] + i[-r_n, r_n]$ of rectangular sets, whose length goes to zero as $n$ goes to infinity, and such that for any $E_n \in R_n$, $d_n$ is such that $E_{n+1} \in R_{n+1}$. $d_n$ is chosen by examining a few digits of $E_n$ and $d_n'$ is chosen by examining a few digits of $E_n'$. These properties allow a simple and fast implementation of the choice of $d_n$.

A. Choice of $d_n$

The diagram presented in Fig. 1 shows the different parameters involved in determining $d_n$. This figure is close to the

Robertson Diagrams that appear in many division algorithms [14]. In the following, we call such a diagram a Robertson diagram.

1) we assume that $E_n$ belongs to the interval $[-s_n, r_n]$, which is the real part of $R_n$.

2) $E_{n+1}$ is equal to $E_n - k \ln \left[ 1 + d_n^2 e^{-2^{n+1}} \right] + \left( d_n^2 + d_n^2 \right) 2^{-2n}$, so the value of $E_{n+1}$ vs. $E_n$ is given by various lines parameterized by $d_n$ and $d_n'$.

3) $d_n$ must be such that for any possible value of $d_n$,

$$E_{n+1} = \left[ -s_{n+1}, r_{n+1} \right]$$

$r_{n+1}$ must be the largest value of $E_{n+1}$ corresponding to the straight line $d_n = 1$ (see Fig 1). That is to say, $r_n$ must satisfy:

$$r_{n+1} = r_n - \ln (1 + 2^{-n})$$

Similarly, the lowest possible value for $E_{n+1}$ must correspond to the value obtained with $d_n = -1 \pm i$. This gives: $s_n = -\frac{1}{2} \sum_{k=1}^{n} \ln (1 + 2^{-k+1}) + 2^{-2k+1}$.

The terms $A_n$, $B_n$, $B_n$, and $B_n$ appearing in the diagram shown in Fig. 1 are equal to:

$$B_n = B_{n+1} + \ln (1 + 2^{-n})$$

$$A_n = A_{n+1} + \frac{1}{2} \ln (1 + 2^{-n+1})$$

$$B_{n+1} = B_{n+1} + \frac{1}{2} \ln (1 + 2^{-n+1})$$

$$B_{n+1} = B_{n+1}$$
One can prove that $\hat{B}_n < \hat{A}_n$, and that $A_n < B_n$. From this, for any $E_n^x \in [-s_n^x, r_n^x]$, the following choices will give a value of $E_{n+1}^x$ between $-s_{n+1}^x$ and $r_{n+1}^x$:

\[
\begin{align*}
\text{if} & \quad E_n^x < -B_n & \text{then} & \quad d_n^x = -1 \\
\text{if} & \quad -B_n \leq E_n^x \leq A_n & \text{then} & \quad d_n^x = -1 \text{ or } 0 \\
\text{if} & \quad A_n \leq E_n^x < B_n & \text{then} & \quad d_n^x = 0 \\
\text{if} & \quad B_n < E_n^x & \text{then} & \quad d_n^x = 1.
\end{align*}
\]

\[ (4) \]

B. Choice of $d_n^y$

We use the relation $E_n^{y+1} = E_n^y - d_n^y \arctan \left( \frac{s_n^y}{1+d_n^{y^2}} \right)$. Fig. 2 shows the Robertson diagram associated to the choice of $d_n^y$. From this, we deduce:

\[ r_n^y = \sum_{k=1}^{\infty} \arctan \left( \frac{s_n^{y^2}}{1+2^{k-2}} \right). \]

The terms $C_n$ and $D_n$ appearing in the diagram are: $C_n = -r_n^y + \arctan \left( \frac{s_n^{y^2}}{1+2^{y-2}} \right)$ and $D_n = r_n^y + 1$. One can prove that $C_n < D_n$. Thus, for any $E_n^y \in [-s_n^y, r_n^y]$, the following choices will give a value of $E_{n+1}^y$ between $-r_{n+1}^y$ and $+r_{n+1}^y$:

\[
\begin{align*}
\text{if} & \quad E_n^y < -D_n & \text{then} & \quad d_n^y = -1 \\
\text{if} & \quad -D_n \leq E_n^y < -C_n & \text{then} & \quad d_n^y = -1 \text{ or } 0 \\
\text{if} & \quad C_n \leq E_n^y < D_n & \text{then} & \quad d_n^y = 0 \\
\text{if} & \quad D_n < E_n^y & \text{then} & \quad d_n^y = 1.
\end{align*}
\]

\[ (5) \]

The convergence domain $R_1$ of the algorithm is:

\[ -0.8298023738 \cdots \leq E_n^x \leq r_n^x = 0.868876517 \cdots \]

\[ -0.749780302 \cdots \leq E_n^y \leq r_n^y = 0.749780302 \cdots \]

C. The Algorithm

Relations (4) and (5) make it possible to find a sequence $d_n$ such that, for $E_1 \in R_1$, $\lim_{n \to \infty} E_n = 0$. Now, let us try to simplify the choice of $d_n$; (4) and (5) involve comparisons that may require the examination of all the digits of the variables, we want to replace these comparisons by the examination of a small number of digits. The parameters $\bar{A} = -1/2$, $A = 1/4$, $C = 3/4$, $p_1 = 3$ and $p_2 = 4$ satisfy for every $n$:

\[
\begin{align*}
2^n B_n & \leq \bar{A} - 2^{-p_1} < \bar{A} \leq 2^n \hat{A}_n \\
2^n A_n & \leq A < A + 2^{-p_1} \leq 2^n B_n \quad (6) \\
2^n C_n & \leq C < C + 2^{-p_2} \leq 2^n D_n.
\end{align*}
\]

Therefore, if we denote $\hat{E}_n^x$ the number obtained by truncating $2^n E_n^x$ after its $p_1$th fractional digit, and $\hat{E}_n^y$ the number obtained by truncating $2^n E_n^y$ after its $p_2$ fractional digit, we obtain, from (4), (5) and (6):

- if $\hat{E}_n^x < \bar{A} - 2^{-p_1}$ then $\hat{E}_n^x < \bar{A}_n$ therefore $d_n^x = -1$ is a valid choice
- if $A < \hat{E}_n^x < \bar{A}$ then $\hat{B}_n \leq E_n^x \leq B_n$ therefore $d_n^x = 0$ is a valid choice
- if $\hat{A} + 2^{-p_1} \leq \hat{E}_n^x$ then $A_n \leq E_n^x$ therefore $d_n^x = 1$ is a valid choice
- if $\hat{E}_n^y < -C - 2^{-p_2}$ then $\hat{E}_n^y < -C_n$ therefore $d_n^y = -1$ is a valid choice
- if $-C \leq \hat{E}_n^y < C$ then $-D_n \leq E_n^y \leq D_n$ therefore $d_n^y = 0$ is a valid choice
- if $C + 2^{-p_2} \leq \hat{E}_n^y$ then $C_n \leq E_n^y$ therefore $d_n^y = 1$ is a valid choice

From this, we deduce the $E$-mode of the BKM algorithm.

**BKM Algorithm—$E$-mode**

- Start with $E_1 \in R_1 = [-0.82980 \cdots, 0.86887 \cdots] + i[-0.74978 \cdots, 0.74978 \cdots]$
- Iterate:

\[
\begin{align*}
E_{n+1} &= E_n - \ln(1 + d_n 2^{-n}) \\
\hat{E}_{n+1} &= \hat{E}_n - \ln(1 + d_n 2^{-n})
\end{align*}
\]

with $d_n = d_n^x + id_n^y$, chosen as follows:

- define $\hat{E}_n$ as the number obtained by truncating the real part of $2^n E_n$ after its 3rd fractional digit, and $\hat{E}_n^y$ as the number obtained by truncating the imaginary part of $2^n E_n$ after its 4th fractional digit.

\[
\begin{align*}
\text{if} & \quad \hat{E}_n^x \leq -\frac{3}{4} & \text{then} & \quad d_n^x = -1 \\
\text{if} & \quad -\frac{1}{4} \leq \hat{E}_n^x < \frac{1}{4} & \text{then} & \quad d_n^x = 0 \\
\text{if} & \quad \hat{E}_n^x \leq \frac{3}{4} & \text{then} & \quad d_n^x = 1 \\
\text{if} & \quad \hat{E}_n^y \leq -\frac{1}{16} & \text{then} & \quad d_n^y = -1 \\
\text{if} & \quad -\frac{1}{16} < \hat{E}_n^y \leq \frac{1}{16} & \text{then} & \quad d_n^y = 0 \\
\text{if} & \quad \hat{E}_n^y > \frac{1}{16} & \text{then} & \quad d_n^y = 1.
\end{align*}
\]

- Result: $\lim_{n \to \infty} L_n = L_1 e^{E_1}$

In practice, instead of computing $E_{n+1} = E_n - \ln(1 + d_n 2^{-n})$ and examining the first digits of $\alpha_n = 2^n E_n$, one would directly compute the sequence $\alpha_{n+1} = 2\alpha_n - 2^{n+1} \ln(1 + d_n 2^{-n})$.

D. Number of Iterations

Now, let us roughly estimate the number of iterations required to obtain a given accuracy. We want to compute $L_1 e^{E_0}$. The sequence $d_n$ defined by the algorithm satisfies $L_1 e^{E_n} = L_1 \prod_{i=0}^{n} (1 + d_i 2^{-i})$. After $n$ iterations of the $E$-mode, we have computed $L_1 \prod_{i=1}^{n} (1 + d_i 2^{-i})$. The relative error made by approximating $L_1 e^{E_n}$ by this value is $|1 - \prod_{i=1}^{n} (1 + d_i 2^{-i})|$, which is bounded by a term equivalent to $2^{-n}$. Therefore, after $n$ iterations of the $E$-mode, we obtain a relative error approximately equal to $2^{-n}$.

1 Since $\bar{A}$, A, and $\bar{E}_n^x$ have at most $p_1$ fractional digits, if $\bar{E}_n^x > \bar{A} - 2^{-p_1}$, then $E_n^x \geq \bar{A}$. 

Fig. 2. The Robertson Diagram for $E_n^x$. 

- The number obtained by truncating $2^n E_n^x$ after its $p_1$ fractional digit, and $\hat{E}_n^y$ the number obtained by truncating $2^n E_n^y$ after its $p_2$ fractional digit, we obtain, from (4), (5) and (6):

- if $\hat{E}_n^x < \bar{A} - 2^{-p_1}$ then $\hat{E}_n^x < \bar{A}_n$ therefore $d_n^x = -1$ is a valid choice
- if $A < \hat{E}_n^x < \bar{A}$ then $\hat{B}_n \leq E_n^x \leq B_n$ therefore $d_n^x = 0$ is a valid choice
- if $\hat{A} + 2^{-p_1} \leq \hat{E}_n^x$ then $A_n \leq E_n^x$ therefore $d_n^x = 1$ is a valid choice
E. Number of Constants Stored

This strategy requires the storage of the constants 
\( \ln(1 + d_k^2 z^{2k+1}) + d_k^2 z^{2k+1} \), where \( d_k^2 = -1, 0, 1 \) and \( \arctan \left( \frac{z}{1 + d_k^2} \right) \). The number of constants to be stored is 9 terms for each value of \( i \). Therefore, we need to store 9n constants. This result can be improved by observing that if \( i > n/2 \), then 
\( \ln(1 + d_k^2 z^{2k+1}) + d_k^2 z^{2k+1} \) and \( \arctan \left( \frac{z}{1 + d_k^2} \right) \) can be replaced by \( d_k^2 z^{2k+1} \) and \( z^{-i} \) with accuracy \( 2^{-n} \). Thus, we only need to store \( 9^n \) constants.

III. COMPUTATION OF THE LOGARITHM FUNCTION (L-MODE)

As shown in the introduction, computing the logarithm of a complex number \( L_1 \) using BKM requires the calculation of a sequence \( d_k, d_k = -1, 0, 1 \) such that \( \lim_{n \to \infty} L_n = 1 \). We deduce:

\[ e_n+1 = 2(e_n + d_k) + d_k e_n 2^{-n+1} \] (7)

If we find a sequence \( e_n \) such that the sequence \( e_n \) is bounded, then \( L_n \) will go to 1. An intuitive solution is to choose \( d_k = -1 \). So, in this section, we consider the following straightforward strategy which consists of building a sequence \( \|e_n\| \leq 3/2 \) as follows:

\[ \|e_n\| \leq 3/2, \text{ this choice will give } d_k \in D. \]

If this algorithm actually gives \( \|e_n\| \leq 3/2 \) for any \( n \), then \( L_n \) will go to 1. From \( \|e_i - e_i| < 2^{-p} \) and \( \|d_k + e_i| \leq 1/2 \), using (7), we deduce:

\[ \|e_{n+1}\| \leq 1 + 2^{-p} + 2^{-n+1}\|d_k e_n\|. \] (8)

The norm \( \| \cdot \| \) satisfies \( \|z^k\| \leq 2\|z\| \cdot 2^k \), therefore:

\[ \|e_{n+1}\| \leq 1 + 2^{-p} + 2^{-n+1}\|e_n\|. \] (9)

If \( n \geq 4, p \geq 4 \), and \( \|e_n\| \leq 3/2 \), then using (9), one can prove that for any \( k \geq n, \|e_n\| \leq 3/2 \). Therefore, if we start the iteration (7) at step 4, from \( e_4 \) satisfying \( \|e_4\| \leq 3/2 \), then the strategy presented above will give a sequence \( d_k \) which fulfills \( \lim_{n \to \infty} L_n = 1 \). This strategy allows computation of the logarithm in a very tiny domain: we can use it to compute the logarithm of \( L_4 \) such that 
\( \|e_4\| = \|16(L_4 - 1)| \leq 3/2 \). So the convergence domain of this algorithm is \( L_4 \in [1 - \frac{3}{32}, 1 + \frac{3}{32}] \).

B. Computation of the Logarithm in a Larger Domain

As in Section III-A, we study the sequence \( e_{k+1} = 2(e_k + d_k) + d_k e_k 2^{-k+1} \), where \( e_k \) is defined as \( 2^k(L_k - 1) \). Our purpose is to start the evaluation at step \( k = 1 \), with \( e_1 \) belonging to a domain \( T \) that will be given later, and to obtain, after a few steps, say \( n \) steps \( (n \geq 3) \), a value \( e_{n+1} \) satisfying \( \|e_{n+1}\| \leq 3/2 \). After this, the strategy presented in the previous section can be used. Our goal is to obtain a convergence domain \( L_1 \in [1/2, 1] \) larger than the previous one. The following algorithm was found through simulations, before being proved.

BKM Algorithm—L-Mode:

- Start with \( L_1 \) belonging to the trapezoid \( T \) delimited by the straight lines \( x = 1/2, x = 1.3, y = x/2, y = -x/2 \). \( T \) is the domain where the convergence is proven, but experimental tests show that the actual convergence domain of the algorithm is larger.

Iterate:

\[ \begin{align*}
L_{n+1} = & \quad L_n(1 + d_k 2^{-n}) \\
E_{n+1} = & \quad E_n(1 + d_k 2^{-n})
\end{align*} \]

with \( d_k = d_k^0 + id_k^1 \), chosen as follows:

- define \( e_k^0 \) and \( e_k^1 \) as the real and imaginary parts of \( e_n = 2^n(L_n - 1) \), and \( e_k^0 \) and \( e_k^1 \) as the values obtained by truncating these numbers after their 4th fractional digits.

- At step 1, we have the equation found at the bottom of the page.

- At step \( n, n \geq 2 \)

\[ \begin{align*}
\text{if } & \quad e_k^0 \leq -1/2 \quad \text{then } d_k^0 = 1 \\
\text{if } & \quad -1/2 < e_k^0 < 1/2 \quad \text{then } d_k^0 = 0 \\
\text{if } & \quad 1/2 \leq e_k^0 \quad \text{then } d_k^0 = -1 \\
\text{if } & \quad c_k^0 \leq -1/2 \quad \text{then } d_k^1 = 1 \\
\text{if } & \quad -1/2 < c_k^0 < 1/2 \quad \text{then } d_k^1 = 0 \\
\text{if } & \quad 1/2 \leq c_k^0 \quad \text{then } d_k^1 = -1
\end{align*} \]

result: \( \lim_{n \to \infty} E_n = E_1 + \ln(L_1) \).

In a practical implementation, instead of computing \( L_n \) and examining the first digits of \( e_n = 2^n(L_n - 1) \), one would directly compute the sequence \( e_n \) using (7).

Proof of the Algorithm: Our goal is to show that if \( L_1 \in T \), then there exists \( n \geq 4 \) such that \( \|e_n\| \leq 3/2 \). Thus the proof of section 3.1 will hold. In order to prove
this, we build a sequence $\beta_k$ of bounding sets, such that for any $L_1 \in T$, $\varepsilon_k \in \beta_k$. Our problem is reduced to show that there exists $n \geq 4$ such that $\beta_n$ is included in the square $||z|| \leq 3/2$. Let us explain how the sequence $\beta_k$ is computed.

The first bounding set $\beta_1$ is equal to $2(T - 1)$. At step $k$, $\beta_k$ is an aggregate of convex polygons, represented by their vertices. A step of the algorithm can be represented by a splitting of the complex plane into 9 convex d-areas. The d-area associated with $\varepsilon \in \{-1, 0, 1, -i, i, -1, i + 1, -i - 1, -i + 1\}$ is the domain $D(\varepsilon)$ such that if $\varepsilon_k \in D(\varepsilon)$, then the algorithm gives $d_k = \varepsilon$. For instance, if $k \geq 2$, then $D(-1 - i)$ is the set of the complex numbers whose real and imaginary parts are greater than 1/2. In $D(\varepsilon)$, the transformation $\varepsilon_{k+1} = 2(\varepsilon_k + \varepsilon) + 3\varepsilon 2^{-k+1}$ is a similarity, i.e., the combination of a rotation and the multiplication by a real factor. A similarity transforms a convex polygon into another convex polygon.

Each convex polygon of $\beta_k$ is split into sub-convex polygons, obtained by intersecting it with the d-areas. Fig. 3 presents the bounding step at step $k$, the various d-areas (for $k \geq 2$), and the splitting of the polygons of $\beta_k$.

Broadly speaking, $\beta_{k+1}$ is obtained by computing the transformation of each sub-convex polygon obtained after the splitting (the image of a polygon is obtained by computing the image of its vertices). However, we have to take into account the fact that the choice of $d_k$ is based on the examination of $\varepsilon_k$ and $\varepsilon_2$, which are obtained by truncating the real and imaginary parts of $\varepsilon_k$ after their 4th fractional digits. For instance, if $\varepsilon_k = \varepsilon_2 + i\varepsilon_1$ belongs to $D(-1)$, this does not prove that $\varepsilon_k$ actually belongs to $D(-1)$. Therefore, to each sub-convex polygon, a "ribbon" of length $2^{-4}$ is added, so that if $\varepsilon_k$ belongs to the "old" subconvex polygon, then $\varepsilon_k$ belongs to the "new" sub-convex polygon. After this, for each "new" sub-polygon, we compute its image by the similarity defined by the value of $d_k$ associated with the polygon (Fig. 4). This gives the new bounding set $\beta_{k+1}$.

The proof by induction that for any $L_1 \in T$, $\varepsilon_k \in \beta_k$ is straightforward. If we find $n \geq 4$ such that all the vertices of the sub-convex polygons of $\beta_n$ are in the square $||z|| \leq 3/2$, then the algorithm is proven. The adequate value of $n$ is 6: this leads to a number of vertices which is much too large to be examined by a paper-and-pencil method. We have used a program for computing all the vertices of $\beta_6$, this program is written in ML, and uses exact rational arithmetic. Fig. 5 shows the bounding sets $\beta_1$, $\beta_4$, and $\beta_6$. Using this program, we have verified that all the vertices of $\beta_6$ are included in the square $||z|| \leq 3/2$.

C. Number of Iterations

Let us estimate the number of iterations required to obtain a given accuracy. The sequence $d_i$ satisfies $\ln(L_1) = -\sum_{k=1}^{\infty} \ln(1 + d_k 2^{-k})$. After $n$ iterations of the L-mode, we have computed $E_1 = \sum_{k=1}^{n} \ln(1 + d_k 2^{-k})$. The absolute error made by approximating $E_1 + \ln(L_1)$ by this value is $\sum_{k=n+1}^{\infty} \ln(1 + d_k 2^{-k})$, which is bounded by a term equivalent to $2^{-n}\sqrt{2}$. From this we deduce that, to obtain an absolute error equal to $2^{-n}$, one needs to perform $n+1$ iterations.

IV. RANGE REDUCTION

The algorithms for computing elementary functions generally converge in some bounded domain. For computing $f(z)$ with an arbitrary value $x$, one usually needs to find a value
$x^*$ belonging to the convergence domain of the algorithm that computes $f$, such that $f(x)$ can be deduced from $f(x^*)$. This operation is called range reduction.

A. Range Reduction for the Complex Exponential Function

We assume that, if $I$ is an interval containing zero, whose length is greater than $p$, we can compute, from any real $x$, an integer $k$ such that $x - kp \in I$. This can be done by performing a few steps of a SRT-like division algorithm. Assume that we want to compute $e^{x + iy}$. BKM allows the evaluation of the exponential function in $R_1 = \{ -0.82980237 \ldots, +0.8687665 \ldots \} + z \{ -0.749780302 \ldots, +0.749780302 \ldots \}$. The range reduction can be performed as follows:

1) Compute $k_y$ such that $y - k_y \frac{\pi}{2}$ belongs to $[-r_1^l, r_1^r]$. Define $y^*$ as $y - k_y \frac{\pi}{2}$.

2) We have: $e^{x + iy} = e^{(k_y \mod 8) \frac{\pi}{2} + iy}$. The multiplication by $e^{(k_y \mod 8) \frac{\pi}{2}}$ looks difficult to reduce to a small amount of additions and shifts. Fortunately, this problem is easily overcome. As an example, let us consider the case $k_y \mod 8 = 1$. The term $e^{\frac{\pi}{2}}$ is equal to $\sqrt{2}(1 + i)$.

A multiplication by this term is avoided by adding $-\frac{1}{2} \ln(2) = \ln x^*$ to $x$, which gives a value $x^*$, and then obtaining: $e^{x + iy} = (1 + i)e^{x^* + iy}$. A multiplication by $1 + i$ is easily reduced to two additions. A similar trick can be used for the other possible values of $k_y \mod 8$.

So, if we define $k_p$ and $\gamma$ as follows:

\[
\begin{align*}
K_0 &= 1 & \text{and} & \gamma_0 &= 0 \\
K_1 &= 1 + i & \text{and} & \gamma_1 &= -\frac{1}{2} \ln(2) \\
K_2 &= i & \text{and} & \gamma_2 &= 0 \\
K_3 &= -1 + i & \text{and} & \gamma_3 &= -\frac{1}{2} \ln(2) \\
K_4 &= -1 & \text{and} & \gamma_4 &= 0 \\
K_5 &= -1 - i & \text{and} & \gamma_5 &= -\frac{1}{2} \ln(2) \\
K_6 &= -i & \text{and} & \gamma_6 &= 0 \\
K_7 &= 1 - i & \text{and} & \gamma_7 &= -\frac{1}{2} \ln(2)
\end{align*}
\]

then, with $p = k_y \mod 8$ and $x^* = x + \gamma_p$, we get $e^{x + iy} = K_pe^{x^* + iy}$.

3) Compute $k_x$ such that $x^* - 2k_x \ln(2)$ belongs to $[-s_1^l, s_1^r]$. Then we obtain $e^{x + iy} = 2^{s_1^l}K_pe^{x^* + iy}$. The exponential of $x^* + iy^*$ is computed using the E-mode of BKM, and the multiplication by $2^{s_1^l}K_p$ is reduced to two additions and a shift.

B. Range Reduction for the Complex Logarithm Function

Let us define the cone $C_p$ as the set of the numbers $z + iy$ such that $|y| \leq x/2$ (see Fig. 6), and $C_1, C_2, C_3, C_4, C_5, C_6,$ and $C_7$ as $C_p \mod 2^k$. For each nonzero element $z$ of $C_0$ there exists an integer $n$ such that $2^nC_0$ belongs to the convergence domain $T$ of the L-mode. Since $C_0 = \lambda C_0$ for any nonnegative real $\lambda$, we easily find:

\[
\begin{align*}
C_1 &= (1 + i)C_0 = \frac{C_0}{-1-i} & C_2 &= (1 - i)C_0 = \frac{C_0}{-1+i} \\
C_3 &= iC_0 = C_p & C_4 &= -iC_0 = \frac{C_0}{1+i} \\
C_5 &= (-1 + i)C_0 = \frac{C_0}{-1-i} & C_6 &= (1 - i)C_0 = \frac{C_0}{1+i} \\
C_7 &= (-1 - i)C_0 = \frac{C_0}{-1+i} & C_8 &= -C_0
\end{align*}
\]

Assume that we want to compute the logarithm of $z_{\text{init}} = x_{\text{init}} + iy_{\text{init}}$. In order to do this, from $z_{\text{init}}$ we want to obtain $z_{\text{Bkm}} = 2^k(d^x + id^y)z_{\text{init}}$, with $d^x, d^y = -1, 0, 1$ and $d^x, d^y$ not simultaneously equal to zero, such that $z_{\text{Bkm}}$ belongs to $T$. After this, $\ln(z_{\text{Bkm}})$ is computed using the L-mode, then $\ln(z_{\text{Bkm}})$ is subtracted from the result (we just need to store the 8 possible values of $\ln(d^x + id^y)$). This reduction is performed in three steps:

1. **Prescaling:** We find $k_1$ such that $z = x + iy = \pm 2^{k_1}z_{\text{init}}$ satisfies $\frac{1}{2} \leq x < 1$.

2. **Search for a Cone $C_p$ Containing $z$:** Now, let us find $p$ such that $z \in C_p$. Since $x > 1/2$, the only possible values of $p$ are $0, 1, 2, 6, 7$. From the definition of the cones $C_p$, we easily find (see Fig. 6):

\[
\begin{align*}
\text{If} & \quad |y| \leq x/2 & \text{then} & \quad z \in C_0 \\
\text{If} & \quad -3x \leq y \leq -x/3 & \text{then} & \quad z \in C_7 \\
\text{If} & \quad x/3 \leq y \leq 3x & \text{then} & \quad z \in C_1 \\
\text{If} & \quad y \leq -2x & \text{then} & \quad z \in C_6 \\
\text{If} & \quad y \geq 2x & \text{then} & \quad z \in C_2
\end{align*}
\]

By taking into account the overlapping of the cones $C_p$, we can replace these comparisons by comparisons involving only a few digits of $x$ and $y$. If we define $\tilde{x}$ and $\tilde{y}$ as the numbers obtained by truncating $x$ and $y$ after their 5th fractional digits (as previously, we assume a binary number system), we easily deduce: $|x - \tilde{x}| \leq 1/32$ and $|y - \tilde{y}| \leq 1/32$. Therefore we have the following:

1. If $|\tilde{y}| \leq \frac{3}{32} - \frac{1}{32}$ then $|y| \leq \frac{1}{2}$, therefore, $z \in C_0$. 

Fig. 6. Second step of the reduction.
to compute the following functions:

2) If \( \frac{\pi}{4} - \frac{\pi}{32} \leq \theta \leq 2\pi + \frac{1}{32} \) then:
   
   a) \( y \geq \frac{\pi}{4} - \frac{\pi}{32} - \frac{\pi}{64} \). Since \( x \geq \frac{1}{2}, \left(\frac{\pi}{4} - \frac{\pi}{32}\right) = \frac{\pi}{8} \geq \frac{\pi}{12} \).
   Therefore, \( y \geq \frac{\pi}{3} \).
   
   b) \( y \leq 2x + \frac{1}{16} + \frac{\pi}{32} \). Since \( x \geq \frac{1}{2}, 3x + (2x - 3x) + \frac{1}{16} + \frac{\pi}{32} \) is less than \( 3x + \frac{1}{16} + \frac{\pi}{32} - \frac{\pi}{2} \), therefore,
   \( y \leq 3x - \frac{\pi}{16} \leq 3x \).

Therefore \( z \in C_1 \).

3) If \(-2\pi - \frac{1}{16} \leq \theta \leq -\frac{\pi}{4} + \frac{1}{32} \), therefore, in a similar fashion, \( z \in C_2 \).

4) If \( y > 2\pi + \frac{1}{16} \) then \( y \geq 2\pi \), therefore \( z \in C_2 \).

5) \( y < -2\pi - \frac{1}{16} \) then, in a similar fashion, \( z \in C_6 \).

Postscaling: We have found a number \( z = x + iy \), an integer \( k_1 \), a sign \( s = \pm 1 \) and a cone \( C_p, p = 0, 1, 2, 6, 7 \) satisfying \( z = s \cdot 2^{k_1} z_{\text{init}}, \frac{1}{2} \leq x \leq 1 \) and \( z \in C_p \). Now, define \( \rho_k \) and \( \ell_k \) as:

\[
\begin{align*}
\rho_0 &= 1 \\
\rho_1 &= 1 + i \\
\rho_2 &= -i \\
\ell_k &= \ln(\rho_k), k = 0, 1, 2, 6, 7.
\end{align*}
\]

The number \( z' = x' + iy' = \rho_k z \) belongs to \( C_0 \). If \( p \neq 0 \), \( z' \) does not necessarily belong to the convergence domain \( T \) of the \( L \)-mode (which is obtained by intersecting \( C_0 \) with the domain \( \{x: x \leq 1\} \)). Define an integer \( k_3 \) such that \( \frac{1}{2} \leq \frac{k_3}{2} \leq 2 \). Then \( \zeta_{\text{Bkm}} = \zeta_{2} z' \in T \). Define \( \ell \) as \( \ell_{\text{p}} \) if \( s = 1 \) and \( \ell_{\text{p}} + \frac{\pi}{2} \) if \( s = -1 \). The logarithm \( L_{\text{Bkm}} \) of \( \zeta_{\text{Bkm}} \) is obtained using the \( L \)-mode, then \( \ell + (k_1 + k_2) \ln(2) \) is subtracted to \( L_{\text{Bkm}} \) in order to obtain the logarithm of \( z_{\text{init}} \). The imaginary part of the final result will be between \(-2\pi \) and \( 2\pi \), so a "correction step"—consisting in adding \( \pm \pi \) or \( \pm 2\pi \)—may be performed if a particular range is required.

V. COMPUTATION OF ELEMENTARY FUNCTIONS

As shown in the previous sections, BKM makes it possible to compute the following functions:

- in \( E \)-mode, \( L_E E_1 \), where \( E_1 \) is a complex number belonging to \( R_1 \),
- in \( L \)-mode, \( E_1 + \ln(L_1) \), where \( L_1 \) belongs to the trapezoid \( T \).

\[
\begin{align*}
\text{Fig. 7. 2-D rotations and arctangents.}
\end{align*}
\]

Therefore, using BKM, one can compute the following functions of \( \pi \).

A. Functions Computable Using One Mode of BKM

1) Real sine and cosine Functions: In the \( E \)-mode of BKM, one can compute the exponential of \( E_1 = 0 + j \theta \), and obtain \( L_1 = \cos \theta + j \sin \theta \pm 2^n \).

2) Real Exponential Function: If \( E_1 \) is a real number belonging to \([-0.8298023738, +0.8688766517]\), the \( E \)-mode will give a value \( L_1 \) equal to \( L_1 \cdot e^{E_1} \pm 2^n \).

3) Real Logarithm: If \( L_1 \) is a real number belonging to \([0.5, 1.3]\), the \( E \)-mode will give a value \( E_o \) equal to \( E_1 + \ln(L_1) \pm 2^n \). Furthermore, in this case, the BKM iteration is reduced to Bragg's algorithm, and it is possible to show that the algorithm gives a correct result if and only if \( L_1 \in \left[\prod_{n=1}^{\infty} (1 + 2^{-n})^{-1}, \prod_{n=1}^{\infty} (1 - 2^{-n})^{-1}\right] \approx [0.419422, 1.73137] \).

4) 2-D Rotations: As pointed out in [11], performing rotations is useful for Fast Fourier Transformation, Digital Filtering, and Matrix Computations. The 2-D vector \( (c \ d) \) obtained by rotating \( (a \ b) \) of an angle \( \theta \) satisfies:

\[
\begin{align*}
c + jd &= (a + jb)e^{j\theta} \quad \text{if} \quad (a \ b) \end{align*}
\]

5) Real arctan Function: From the relation:

\[
\ln(x + iy) = \begin{cases} 
\frac{1}{2} \ln(x^2 + y^2) + i \arctan \left( \frac{y}{x} \right) & \text{if } x > 0 \\
\frac{1}{2} \ln(x^2 + y^2) + i \left( \pi + \arctan \left( \frac{y}{x} \right) \right) & \text{if } x < 0 
\end{cases}
\]

one can easily deduce that, if \( x + iy \) belongs to the convergence domain of the \( L \)-mode of BKM, then \( \arctan y/x \) is the limit value of the imaginary part of \( E_o \), assuming that the \( L \)-mode is used with \( E_1 = 0 \) and \( L_1 = x + iy \) (see Fig. 7).
B. Functions Computable Using Two Consecutive Modes of BKM

Using two BKM operations, one can compute many functions. Some of these functions are the following.

1) Complex Multiplication and Division: The product \( z \cdot e^{i \theta} \) is evaluated as \( z \cdot e^{i \theta} \), while \( z / t \) is evaluated as \( z \cdot e^{-i \theta} \) (see Fig. 8). One can compute \((ab) e^z \) or \((\frac{a}{b}) e^z\), where \( a, b \) and \( z \) are complex numbers, using the same operator, by choosing \( E^0 \) equal to the real part of \( z \), and \( E^i \) equal to the imaginary part of \( z \).

2) Computation of \( x \sqrt{a} \) and \( y \sqrt{a} \) in Parallel (\( x \), \( y \), and \( a \) are real numbers): we use the relation \( \sqrt{a} = e^{\frac{1}{2} \ln(a)} \), see Fig. 9. One can also compute \( x \sqrt{a} \) and \( y \sqrt{a} \).

3) Computation of Lengths and Normalization of 2-D Vectors: As shown previously, the L-mode allows the computation of \( F = \frac{1}{2} \ln(a^2 + b^2) \) and \( E = \ln(a^2 + b^2) \), where \( a \) and \( b \) are real numbers. Using the E-mode, we can compute \( e^E \) or \( e^{-F} \) (See Fig. 10).

VI. COMPARISON WITH CORDIC

In order to obtain \( p \) significant bits, CORDIC and BKM roughly need \( p \) iterations. BKM requires more hardware than CORDIC: BKM needs the storage of \( \frac{p}{2} \) constants, while CORDIC needs the storage of \( p \) constants to compute trigonometric and hyperbolic functions. Since these constants are represented by \( p \) digits, both algorithms need a \( O(p^2) \) area for storage of the constants. Both algorithms need a shifter able to perform an \( n \)-position shift at step \( n \). A barrel shifter makes it possible to perform a \( n \)-position shift (for any \( n \leq p \)) in constant time, and lies in an area \( O(n^2) \). Since the area complexity of most adders is better than \( O(n^2) \), the area complexity of CORDIC and BKM is \( O(n^2) \). The computations performed during a BKM iteration are:

- For the variable:
  \[ E_n = 2a_n - 2n \ln\left[1 + d_n^2 e^{-n+1} + (a_n^2 + d_n^2) 2^{-2n}\right] \]
  if, instead of computing \( E_n \) and examining the first digits of \( a_n = 2^n E_n \), we directly compute \( a_n \).

- For the variable:
  \[ L_n = 2(c_n + d_n) + (d_n e^n - d_n e^n) 2^{-n+1} \]
  if, instead of computing \( L_n \) and examining the first digits of \( c_n = 2^n (L_n - 1) \), we directly compute \( c_n \).

So BKM looks more complicated than CORDIC. As a matter of fact, in order to compare CORDIC and BKM, we have to assume that we use a redundant number system. Using such a system, the time complexities of both algorithms are \( O(p) \). As pointed out in many papers dealing with CORDIC, efficient use of CORDIC with such a number system requires a doubling of the iterations in space [7] or in time [16], [1]. Doubling the CORDIC iterations in time gives:

\[
\begin{align*}
x_{n+1} &= x_n - d_n y_n 2^{-n} - d_n^2 x_n 2^{-2n-2} \\
y_{n+1} &= y_n + d_n x_n 2^{-n} - d_n^2 x_n 2^{-2n-2} \\
z_{n+1} &= z_n - 2d_n \arctan 2^{-n-1}. 
\end{align*}
\]

This iteration is at least as complex as the BKM iteration (at step \( n \) one needs to perform an \( n \)-position shift and a \( 2n - 2 \)-position shift: this requires a larger shifter, or several consecutive shifts). Doubling the iterations in space requires more control: in the branch CORDIC method [7], one needs to compare at each step the values given by two CORDIC modules. Furthermore, doubling iterations makes it possible to obtain a constant scale factor, but this scale factor remains different from 1, therefore, for computing many functions, one needs to perform a multiplication after the CORDIC iterations. So, although both methods have the same time and space complexities, BKM looks more interesting when using a redundant number system.

VII. CONCLUSION

We have proposed a new algorithm for the computation of many elementary functions (complex exponential and logarithm functions, complex multiplication, complex functions \((ab)e^z\) and \((\frac{a}{b})e^z\), real functions sin, cos, \arctan z, \ln(x^2 + y^2), x\sqrt{a^2 + b^2}, x/\sqrt{a^2 + b^2}, \) and 2-D rotations). This algorithm matches the CORDIC algorithm, since it allows the use of a redundant number system without any scale factor problem and allows the computation of more functions.

ACKNOWLEDGMENT

We thank C. Paulin for her precious knowledge of ML and her careful reading of the manuscript.
REFERENCES


Jean-Claude Bajard was born in Saint-Etienne, France, in 1957. He received the Ph.D. degree in computer science in 1993 from the Ecole Normale Superieure de Lyon. He taught mathematics in high school from 1979 to 1990, and served as an assistant professor at Ecole Normale Superieure de Lyon in 1993. He joined the Lab. LIM, Universite de Provence, Marseille, France in 1993.

Dr. Bajard's research interests include computer arithmetic and VLSI design.

Sylvanus Kla was born in Ivory Coast in 1963. He received the Engineer degree in civil engineering from Ecole Nationale Superieure des Travaux Publics, Yamoussoukro, Ivory Coast, in 1988, the M.S. degree in applied mathematics from Grenoble University, France, in 1989, and the Ph.D. degree from Lyon University, France, in 1993.

His research interests are in computer arithmetic and computer architecture.

Jean-Michel Muller (M'87) was born in Grenoble, France, in 1961. He received the Engineer degree in applied mathematics and computer science in 1983, and the Ph.D. in computer science in 1985, both from the Institut National Polytechnique de Grenoble, France.

He was posted from 1986 to 1989 with Tim3-Imag Laboratory, Grenoble. He is currently with CNRS, Lab. LIP, Ecole Normale Superieure de Lyon, Lyon. He teaches computer arithmetic in the Ecole Normale Superieure de Lyon. His research interests include computer arithmetic and computer architecture.

Dr. Muller served as General Chairman of the 10th Symposium on Computer Arithmetic (Grenoble, France, June, 1991).