APPLICATION OF MODULAR COMPUTING TECHNOLOGY TO NUMBER NORMALIZATION IN FLOATING POINT ARITHMETIC

MIKHAIL SELIANINAU

ABSTRACT

In the present paper, we deal with the methodology of mantissa normalization on the basis of parallel algorithmic structures of modular arithmetic. The use of interval-modular form and basic integral characteristics of modular code is fundamental for construction of floating-point modular computing arithmetic. The proposed method of mantissa normalization in the minimal redundant modular number system is based on the parallel algorithm of multiplication by constant with overflow check.

1. Introduction

The problem of providing a high speed and sufficient accuracy of calculations has always been the actual direction of theoretical research in the field of computer arithmetic. At the present time, this problem gets more and more clear-cut applied aspect which appears first of all as a solution of engineering and scientific problems of high dimensionality and high time complexity. Most numerical methods operate over the field of real numbers, hence, the problem of creating new ways of their computer approximation and the organization of parallel processing is actual.

Representation of numbers in the floating point form is used in computer systems that serve to solve a wide range of scientific and technical problems. This is caused by the fact that the use of floating point form allows us to expand considerably the range of representable numbers in comparison with the fixed point while the high precision of representation is required.

The research directed to the development of new methods and algorithms of high-speed and high-precision floating point calculations are of great importance as they allow us to perform effective parallelization of calculations.
and to minimize the dependence of the speed of arithmetic operations execution on the computing accuracy of these operations.

One of the most promising lines of research in the field of organization of high-speed high-precision calculations is the implementation of nonconventional ways of information encoding and corresponding variants of computer arithmetic. Numerical systems with parallel structure and, first of all, the modular number systems (MNS) which are characterized by the maximum level of internal parallelism play a significant role in the development of the approach outlined above [1–4]. The method of implementation of high-precision arithmetic for the processing of numbers of big digit capacity with the fixed and floating point in the MNS basis is most consistent with the development vector of modern high-performance computing technologies and systems.

2. Minimal redundant modular coding

Currently, the research aimed to optimizing modular computing structures (MCS) is intensively conducted in parallel with extending the sphere of applications of modular arithmetic (MA). These structures represent the unique means of computation processes decomposition into independent elementary subprocesses defined on mathematical models with elements of small digit capacity [1–6].

Using of modular redundant coding is highly efficient in study of MCS. In this connection, the modular codes (MC) having minimum redundancy, i.e. the codes of minimal redundant modular number systems (MRMNS), are of particular interest. The basis of the principle of minimal redundancy coding consist in the use of some less powerful operating ranges in comparison with the ranges determined by the Chinese remainder theorem [1, 2, 5, 6].

In the set $\mathbb{Z}$ of real integer numbers, an MRMNS is determined by pairwise prime natural modules $m_1, m_2, \ldots, m_k$ ($m_k \geq 2m_0 + k - 2$, $m_0$ is the additional real module which satisfies the condition $m_0 \geq k - 2$, $k \geq 2$) by defining a homomorphic mapping

$$
\phi: \mathbb{D} \rightarrow \mathbb{Z}_{m_1} \times \mathbb{Z}_{m_2} \times \ldots \times \mathbb{Z}_{m_k},
$$

$$(\mathbb{Z}_{m_i} = \{0, 1, \ldots, m_i - 1\}, i = 1, 2, \ldots, k)$$

on the working range

$$
\mathbb{D} = \mathbb{Z}_{2M}^- = \{-M, -M + 1, \ldots, M - 1\},
$$

$$(M = m_0M_{k-1}, M_{k-1} = \prod_{l=1}^{k-1} m_l).$$
In this case, the vector
\[(\chi_1, \chi_2, \ldots, \chi_k) \in \mathbb{Z}_{m_1} \times \mathbb{Z}_{m_2} \times \ldots \times \mathbb{Z}_{m_k},\]
is assigned to each \(X \in \mathcal{D}\), where \(\chi_1 = |X|_{m_1}, \chi_2 = |X|_{m_2}, \ldots, \chi_k = |X|_{m_k}\) are the residues of division of an integer \(X\) by natural modules \(m_1, m_2, \ldots, m_k\); the notation \(|x|_m\) is used to designate an element of the residue ring \(\mathbb{Z}_m = \{0, 1, \ldots, m-1\}\) which is congruent to the rational value \(x\) modulo \(m\).

Similar to a conventional MNS with the bases \(m_1, m_2, \ldots, m_k\), in an MRMNS all the ring operations over any two integer numbers \(A\) and \(B\) represented by their MC:
\[A = (\alpha_1, \alpha_2, \ldots, \alpha_k), B = (\beta_1, \beta_2, \ldots, \beta_k),\]
\(\alpha_i = |A|_{m_i}, \beta_i = |B|_{m_i}; i = 1, 2, \ldots, k\)
are also performed independently for each module, i.e. according to the rule
\[A \circ B = (\alpha_1, \alpha_2, \ldots, \alpha_k) \circ (\beta_1, \beta_2, \ldots, \beta_k) = \]
\[(|\alpha_1 \circ \beta_1|_{m_1}, |\alpha_2 \circ \beta_2|_{m_2}, \ldots, |\alpha_k \circ \beta_k|_{m_k}); \circ \in \{+ , - , \times\}. \quad (1)\]

The main advantage of MA over the arithmetic of positional number systems consists just in the property (1). The absence of carry between the adjacent digits of numbers represented in an MNS allows us to perform the modular arithmetic operations easy and fast. Since the components of the MC have a small code length and the ring operations in the MNS are performed independently for each module, then the MA gives essentially new possibilities to increase the computation speed.

Computational complexity, efficiency and realizable properties of one or another variant of MA are primarily dependent on the selected set of basic integral characteristics of MC (ICMC). These characteristics which values are determined by the digits of MC allow us to estimate the magnitude of the corresponding numbers or to receive their full positional codes [1, 2].

With respect to the mapping \(\phi\) which defines an MRMNS with the bases \(m_1, m_2, \ldots, m_k\) and operating range \(\mathcal{D}\), there is an inverse mapping \(\phi^{-1}\) which establishes a correspondence between the vectors \((\chi_1, \chi_2, \ldots, \chi_k)\) and the numbers \(X \in \mathcal{D}\) by the rule
\[X = \sum_{i=1}^{k-1} M_{i,k-1} \chi_{i,k-1} + I(X)M_{k-1}, \quad (2)\]
where \(\chi_{i,k-1} = |M_{i,k-1}^{-1}|_{m_i}\), \(M_{i,k-1} = M_{k-1}/m_i\), \(I(X)\) is an interval index (II) of the number \(X\). The expression (2) is called an interval-modular form (IMF) of the number \(X\) [1, 2, 5].
The triviality of calculation of II is a key element of the principle of minimum redundancy modular coding. The value of the II is uniquely determined in the MRMNS by its residue modulo $m_k$ and is reduced to the calculation of a linear combination of low-bit remainders [2, 5].

In the MRMNS with the modules $m_1, m_2, \ldots, m_k$ ($m_k \geq 2m_0 + k - 2$, $m_0 \geq k - 2$), along with the integer numbers the fractions of the form $x = X/M$, where $X \in \mathbb{D}$, can also be considered. The set of fractions mentioned above forms some finite model of real numbers from the interval $(-1, 1)$. In the MNS, a fraction of a given type is generally defined by the MC $(\chi_1, \chi_2, \ldots, \chi_k)$ of the number $X$ ($\chi_i = |X|_{m_i}; \ i = 1, 2, \ldots, k$).

Using the IMF (2) of real integer numbers and the calculation procedure of an II as a basis for all algorithmic construction provides simplicity of implementation of decoding mapping in MRMNS and consequently of other non-modular operations. This leads to a significant improvement of the arithmetic properties of the MNS.

3. Overflow check and sign detection in the MRMNS

The interval-index characteristic $I(X)$ allows us to check quite simply whether or not the result number $X$ belongs to the operating range under consideration as well as to identify the negative and non-negative area of the specified range, i.e. to solve the problem of overflow and sign detection.

In the development of computer arithmetic of the MRMNS, along with the II $I(X)$, a minimal ICMC $\Theta(X)$ corresponding to the number $X$ in the system with modules $m_1, m_2, \ldots, m_{k-1}, m_0$ determined by the relation

$$|X|_M = \sum_{i=1}^{k-1} M_{i,k-1}X_{i,k-1} + \hat{I}(X)M_{k-1} - \Theta(X)M$$

plays a key role [1]. Here $\hat{I}(X) = |I(X)|_{m_0}$ is a computer II of the number $X$.

We introduce the following definition. **Definition 1.** An integer number

$$J(X) = \left\lfloor \frac{I(X)}{m_0} \right\rfloor$$

is called the main II of an arbitrary integer number $X$ with respect to the module $m_0$ (the integer part of a real number $x$ is denoted by $\lfloor x \rfloor$).

**Definition 2.** An integer number

$$N(X) = \left\lfloor \frac{X}{M} \right\rfloor$$

(5)
is called an interval number of an arbitrary integer number $X$ with respect to the modules $m_1, m_2, \ldots, m_{k-1}, m_0$.

By applying the Euclidean lemma from the theory of divisibility [7] to an $I(X)$ in the form $I(X) = I(X) + J(X)m_0$, we have from (2)

$$X = \sum_{i=1}^{k-1} M_{i,k-1} \chi_{i,k-1} + I(X)M_{k-1} + J(X)M.$$

Subtracting and adding the value $\Theta(X)M$ in the right-hand side of this equality and considering (3), we receive

$$X = \sum_{i=1}^{k-1} M_{i,k-1} \chi_{i,k-1} + I(X)M_{k-1} - \Theta(X)M + J(X)M + \Theta(X)M =

= |X|_M + (J(X) + \Theta(X))M. \quad (6)$$

Hence, according to the Euclidean lemma, we conclude that for an interval numbers $N(X)$ (see (5)) the following relation is true

$$N(X) = J(X) + \Theta(X). \quad (7)$$

As it follows from (6) and (7), a necessary and sufficient condition for the number $X$ to belong to the considered range is an execution of one of the equalities: $J(X) + \Theta(X) = -1$ or $J(X) + \Theta(X) = 0$. In other words, in the MRMNS a control of overrange of an integer number $X$ is reduced to forming a flag

$$\Omega = \begin{cases} 0 & \text{if } J(X) + \Theta(X) \in \{-1, 0\}, \\ 1 & \text{in other cases.} \end{cases} \quad (8)$$

The obtained relations (6) and (7) also allow us to determine the sign $S(X)$ of a number $X$:

$$S(X) = \begin{cases} 0 & \text{if } J(X) + \Theta(X) \geq 0, \\ 1 & \text{if } J(X) + \Theta(X) < 0. \end{cases} \quad (9)$$

Thus, in the MRMNS the operations of overflow check (8) and of sign detection (9) are actually reduced to calculation of the main $I(X)$ and the minimal ICMC $\Theta(X)$ corresponding to the integer number $X$ in the auxiliary MNS with the modules $m_1, m_2, \ldots, m_{k-1}, m_0$.

In [1] it was proved that if the module $m_0$ satisfies the condition $m_0 \geq k - 2$, then the minimal ICMC $\Theta(X)$ possesses only two values: 0 or 1.

Let us note that in the MRMNS for implementation of the most commonly used arithmetic operations (addition, subtraction and multiplication of numbers with overflow check), it is best to use the strictly symmetric
range \( D_0 = \{-M + 1, -M + 2, \ldots, M - 1\} \) which originates from the operating range \( D \) by deleting a point \(-M\). This feature introduces some changes in the process of forming an overflow flag \( \Omega \) because now the overflow takes place when \( N(X) \in \{-1, 0\} \) and \( X \neq -M \). Thus, when using the symmetric range, it is necessary to substitute the relation (8) by the following modification:

\[
\Omega = \begin{cases} 
0 & \text{if } J(X) + \theta(X) \in \{-1, 0\} \text{ and } X \neq -M, \\
1 & \text{in other cases.}
\end{cases}
\] (10)

4. Features of floating point numerical data representation

It is known that any real number \( x \) in a positional number system can be represented in exponential form: \( x = m(x) p^{n(x)} \), where \( m(x) \) is the mantissa, \( p \) is the base radix and \( n(x) \) is the number exponent.

For representation of floating point data, the number record in a normalized form is usually used. In this case, the mantissa \( m(x) \) satisfies the following conditions: it must be a proper fraction and have the nonzero digit in the first number position after the floating point. It should be emphasized that the requirement of normalization of numbers is introduced to ensure the maximum accuracy of their representation.

As a result of calculations, the normalization is often violated, so it is necessary to restore it. For example, when performing arithmetic operations of addition and subtraction of mantissas the normalization of a resultant mantissa can be violated at one digit to the left or at arbitrary number of digits to the right.

If normalization is violated to the left, then the normalization of the result is executed by shifting the resultant mantissa at one digit to the right and the increment of the number exponent by one.

If normalization is violated to the right, then the normalization of the result is executed by shifting the resultant mantissa at one digit to the left and the decrement of the number exponent by one. These steps should be continued till the condition \( 1/p \leq |m(x)| < 1 \) is met.

5. Mantissa normalization in the MRMSS

On the basis of minimal redundant modular algorithmic structures it is possible to construct different variants of MA of real numbers not only with a fixed point but also with a floating point. For this purpose it is necessary to develop additionally fast methods and algorithms of normalization of mantissas represented in the minimal redundant MC.

The operation of number normalization refers to the category of non-modular operations. The main methods of mantissa normalization in the
M RMNS can be synthesized on the basis of parallel algorithms of forming
different ICMC (for example, coefficients of a symmetric polyadic code),
high-speed scaling procedures and arithmetic operations with overflow check [2, 8].

Let us consider the approach to constructing the normalization proce-
du res when normalization is violated to the right which is based on the
multiplication by constants with overflow check.

In the MRMNS with the modules \( m_1, m_2, \ldots, m_k \) \( (m_k \geq 2m_0 + k - 2, \ m_0 \geq k - 2) \), a mantissa \( m(x) \) is represented in terms of a fraction
\( m(x) = X/M \) which is defined by the MC \( (\chi_1, \chi_2, \ldots, \chi_k) \) of a number \( X \) \( (\chi_i = \lvert X \rvert_{m_i}, i = 1, 2, \ldots, k) \). As the normalization is violated to the
right \( 0 \leq \lvert m(x) \rvert < 1/p \), then \( 0 \leq \lvert X \rvert < M/p \). Therefore, the mantissa
normalization is reduced to transformation of a number \( X \).

To normalize a number \( X \) in MRMNS (\( X \) is an element of operating range
of the MRMNS, \( X \neq 0 \)) means to find its representation in the exponential
form

\[
X = m(X)p^{-n(X)},
\]

(11)

where \( m(X) \) is an integer-valued mantissa satisfying the condition \( M/p \leq \lvert m(X) \rvert < M \); \( n(X) \) is an exponent of a number \( X \) and \( p \) is a radix of the
exponential representation.

The relation (11) indicates that normalization of a number \( X \) is actually reduced to computation of the exponent \( n(X) \). The value \( n(X) \) is uniquely
determined by the condition of normalization of the mantissa \( m(X) \) written as

\[
M/p \leq \lvert X \rvert p^{n(X)} < M.
\]

From the relationship (11) it follows that the exponent \( n(X) \) of the initial
number \( X \) coincides with the number \( \nu \) of the penultimate element of a sequence
\( X^{(0)} = X, X^{(1)} = X^{(0)} p, X^{(2)} = X^{(1)} p, \ldots, X^{(\nu)} = X^{(\nu - 1)} p, \)
\( X^{(\nu+1)} = X^{(\nu)} p (\nu \geq 0) \), where numbers \( X^{(0)}, X^{(1)}, \ldots, X^{(\nu)} \) belong and
\( X^{(\nu+1)} \) does not belong to the range of MRMNS.

Therefore, to obtain the value \( n(X) = \nu \) it is enough to generate flags
\( \Omega(X^{(j)}) \) which indicate that the numbers \( X^{(j)} = X p^j \ (j = 1, 2, \ldots, \nu) \)
belong and the number \( X^{(\nu+1)} \) does not belong to the number system range
(see (8), (10)).

Let us introduce the notation

\[
\chi_{i,k-1}^{(l)} = \left\lfloor \frac{p^l M_{i,k-1}^{(l)} X^{(l)}}{m_i} \right\rfloor_{m_i} = \left\lfloor \frac{p^l M_{i,k-1}^{(l)} X^{(l)}}{m_i} \right\rfloor_{m_i} = \left\lfloor \frac{p^l \chi_{i,k-1}^{(l)}}{m_i} \right\rfloor_{m_i} \tag{12}
\]
(i = 1, 2, . . . , k−1; l = 0, 1, . . . , ν + 1). Using the IMF (2) and the Euclidean lemma, we transform a number \( X^{(j)} (j = 1, 2, . . . , ν + 1) \) as follows:

\[
X^{(j)} = pX^{(j-1)} = p \left( \sum_{i=1}^{k-1} M_{i,k-1} X^{(j-2)} + I(X^{(j-1)}) M_{k-1} \right) =
\]

\[
= \sum_{i=1}^{k-1} M_{i,k-1} \left( pX^{(j-2)} + \left[ pX^{(j-1)} / m_i \right] m_i \right) + pI(X^{(j-1)}) M_{k-1} =
\]

\[
= \sum_{i=1}^{k-1} M_{i,k-1} X^{(j-1)} + \left( pI(X^{(j-1)}) + \sum_{i=1}^{k-1} \left[ pX^{(j-1)} / m_i \right] \right) M_{k-1}.
\]

Hence, we conclude that for an II of a number \( X^{(j)} \) the following formula is true

\[
I(X^{(j)}) = pI(X^{(j-1)}) + \sum_{i=1}^{k-1} \left[ pX^{(j-1)} / m_i \right]. \tag{13}
\]

Therewith, the main II \( J(X^{(j)}) = \lfloor I(X^{(j)}) / m_0 \rfloor \) of a number \( X \) with respect to the modules \( m_1, m_2, . . . , m_{k-1}, m_0 \) is determined by the relation

\[
J(X^{(j)}) = \left\lfloor \left( pI(X^{(j-1)}) + \sum_{i=1}^{k-1} \left[ pX^{(j-1)} / m_i \right] \right) / m_0 \right\rfloor. \tag{14}
\]

To establish whether the number \( X^{(j)} \) is an element of the number system range it is enough to calculate the corresponding minimal ICMC \( \Theta(X^{(j)}) \) in the MNS with modules \( m_1, m_2, . . . , m_{k-1}, m_0 \) by means of the ICMC generation algorithm [2, 8] and further to determine a flag \( \Omega(X^{(j)}) \) by applying one of the formulas (8) or (10). The unit value of a flag \( \Omega(X^{(j)}) \) specifies that the number \( X^{(j)} \) exceeds the limits of operating range.

The algorithm of normalization of a number \( X = (\chi_1, \chi_2, . . . , \chi_k) (X \neq 0) \) based on the assumptions stated above is reduced to execution of the following steps.

N.1. The computer II \( \hat{I}_k(X) \) of the initial number \( X \) is computed in accordance with the calculated relations [2, 5]

\[
\hat{I}_k(X) = \left\lfloor \sum_{i=1}^{k} R_{i,k}(\chi_i) \right\rfloor_{m_k};
\]

\[
R_{i,k}(\chi_i) = \left\lfloor \frac{\chi_i M_{k-1}}{m_k} \right\rfloor (i \neq k) \quad R_{k,k}(\chi_k) = \left\lfloor \frac{\chi_k}{M_{k-1}} \right\rfloor_{m_k}.
\]
Then the characteristics \( \hat{I}(X) \) and \( J(X) \) are determined by the tabular method

\[
\hat{I}(X) = \begin{cases} 
\lfloor \hat{I}_k(X) \rfloor_{m_0} & \text{if } \hat{I}_k(X) < m_0; \\
\lfloor \hat{I}_k(X) - m_k \rfloor_{m_0} & \text{if } \hat{I}_k(X) \geq m_k - m_0 - k + 2;
\end{cases}
\]

\[
J(X) = \begin{cases} 
\lfloor (\hat{I}_k(X)/m_0) \rfloor & \text{if } \hat{I}_k(X) < m_0; \\
\lfloor (\hat{I}_k(X) - m_k)/m_0 \rfloor & \text{if } \hat{I}_k(X) \geq m_k - m_0 - k + 2.
\end{cases}
\]

N.2. Let us assume that \( j = 1, X^{(0)} = X \) and \( \chi_i^{(0)} = \chi_i \) (\( i = 1, 2, \ldots, k \)).

N.3. The MC \( (\chi_1^{(j)}, \chi_2^{(j)}, \ldots, \chi_k^{(j)}) \) of the number \( X^{(j)} = pX^{(j-1)} \) (\( \chi_i^{(j)} = |p\chi_i^{(j-1)}|_{m_i}; \ i = 1, 2, \ldots, k \)) and its interval-index characteristics

\[
\hat{I}(X^{(j)}) = \left| R_0(\hat{I}(X^{(j-1)})) + \sum_{i=1}^{k-1} R_i(\chi_i^{(j-1)}) \right|_{m_0} \tag{15}
\]

and

\[
J(X^{(j)}) = pJ(X^{(j-1)}) + \left| p\hat{I}(X^{(j-1)})/m_0 \right| + \\
\left| \left( R_0(\hat{I}(X^{(j-1)})) + \sum_{i=1}^{k-1} R_i(\chi_i^{(j-1)}) \right) /m_0 \right| \tag{16}
\]

where

\[
R_0(\hat{I}(X^{(j-1)})) = \left| p\hat{I}(X^{(j-1)}) \right|_{m_0};
\]

\[
R_i(\chi_i^{(j-1)}) = \left| p M_i^{(j-1)} \chi_i^{(j-1)} \right|_{m_i} = \left[ p\chi_i^{(j-1)}/m_i \right],
\]

are formed using the digits of the MC \( (\chi_1^{(j-1)}, \chi_2^{(j-1)}, \ldots, \chi_k^{(j-1)}) \) of the number \( X^{(j-1)} \) and the values \( \hat{I}(X^{(j-1)}) \) and \( J(X^{(j-1)}) \).

The calculating relations (15) and (16) follow directly from (13) and (14) if one takes into account the designation (12) and considers that

\[
I(X^{(j-1)}) = \hat{I}(X^{(j-1)}) + J(X^{(j-1)}) m_0
\]

and

\[
p\hat{I}(X^{(j-1)}) = \left| p\hat{I}(X^{(j-1)}) \right|_{m_0} + \left| p\hat{I}(X^{(j-1)})/m_0 \right| m_0.
\]

N.4. The minimal ICMC \( \Theta(X^{(j)}) \) corresponding to the number \( X^{(j)} \) in the MNS with modules \( m_1, m_2, \ldots, m_{k-1}, m_0 \) is computed using the ICMC
generation algorithm [2, 8]. As the input data of algorithm we specify the values \( \chi^{(j)}_1, \chi^{(j)}_2, \ldots, \chi^{(j)}_{k-1} \) and \( \hat{I}(X^{(j)}) \).

N.5. According to the obtained values of characteristics \( J(X^{(j)}) \) and \( \Theta(X^{(j)}) \) the flag \( \Omega(X^{(j)}) \) is formed by applying one of the formulas (8) or (10). If the number \( X^{(j)} \) is an element of a system range \( \Omega(X^{(j)}) = 0 \), then \( j \) is incremented by one \( (j = j + 1) \) and whereupon we move to the step N.3, otherwise \( \Omega(X^{(j)}) = 1 \) the cyclic process of multiplication by a constant \( p \) is finished. Let it take place in the case \( j = \nu + 1 \ (\nu \geq 0) \). Then the values \( X^{(\nu)} \) and \( \nu \) are fixed as the required values of the mantissa \( m(X) \) and the exponent \( n(X) \) of a number \( X \). Hence, the operation of normalization is completed.

The normalization algorithm N.1 - N.5 becomes the simplest one for \( p = 2 \). In this case the values \[ px^{(j-1)}/m_i = 2x^{(j-1)}/m_i \] \( (i = 1, 2, \ldots, k-1) \) can be equal only to 0 or 1. Therefore, after the computation of the II \( I(X) \) of the initial number \( X \) the calculating relations (13)–(16) for the interval-index characteristics of a number \( X^{(j)} \) for each \( j \geq 1 \) may be implemented at a single module clock cycle, for example, by the tabular method.

References

Mikhail Selianinou
JAN DŁUGOSZ UNIVERSITY,
INSTITUTE OF TECHNICAL EDUCATION AND SAFETY,
42-200 CZĘSTOCHOWA, AL. ARMII KRJOWEI 13/15, POLAND
E-mail address: m.selianinov@ajd.czest.pl