Digital Filter Architecture with Calculations in the Residue Number System by Winograd Method $F(2 \times 2, 2 \times 2)$

Maria Valueva¹, Graduate Student Member, IEEE, Pavel Lyakhov¹, Georgii Valuev², Nikolay Nagornov³

¹North-Caucasus Center for Mathematical Research, North-Caucasus Federal University, 355017 Stavropol, Russia
²Department of Mathematical Modeling, North-Caucasus Federal University, 355017 Stavropol, Russia
³Department of Automation and Control Processes, Saint Petersburg Electrotechnical University "LETI", 197376 Saint Petersburg, Russia

Corresponding author: Maria Valueva (mriya.valueva@mail.ru).

Pavel Lyakhov and Maria Valueva were supported by the Russian Foundation for Basic Research (Project No. 19-07-00130 A) and Council for Grants of the President of the Russian Federation (Projects No. SP-126.2019.5 and No. MK-3918.2021.1.6).

ABSTRACT Improving the technical characteristics of digital signal processing devices is an important problem in many practical tasks. According to the Winograd method, the paper proposes the architecture of a device for two-dimensional filtering in a residue number system (RNS) with moduli of a special type. The work carried out the technical parameters theoretical analysis of the proposed filter architecture for different RNS moduli sets by the "unit-gate"-model. In addition, the proposed architecture is compared with known digital filter implementations. The theoretical analysis results showed that the proposed filter architecture makes it possible to increase the signal processing speed by 1.33 – 6.90 times, compared with the known device implementations. Also, in the paper, the hardware simulation of the proposed filter architecture was performed on FPGA, which showed that the performance of the proposed device is 1.31 – 4.12 times higher than known digital filter architectures. The research results can be used in digital signal processing systems to increase their performance and reduce hardware costs. In addition, the developed architectures can be applied in the development of hardware accelerators for complex digital signals analysis systems.

INDEX TERMS Digital Filters, Residue Number System, Winograd Method

I. INTRODUCTION Digital signal filtering is widely applied in various areas such as medicine [1, 2], geolocation [3], video surveillance systems [4], quality control in production [5], and many others. Performance plays a central role in these practical tasks. Hardware implementation of digital filtering allows increasing the speed of signal processing systems [6]. Therefore, improving digital filter technical characteristics is a significant challenge.

The main computational load during filtering consists in multiply performing the multiplication operation. One of the approaches to increasing a digital filter’s speed is to reduce the number of multiplications. The paper [7] proposes the Winograd filtering method, which reduces the number of multiplications in the filtering process by increasing the number of additions. The authors of [8] presented a software implementation of the Winograd method and applied it in a convolutional layer of a neural network with calculations on a graphical processor. In [9], the authors developed a hardware accelerator on Field-Programmable Gate Array (FPGA) based on the Winograd method for the convolutional layer of the neural network.

Another approach to increase the speed of devices is parallel computations. The residue number system (RNS) is a non-positional number system, which performs numbers as small residues modulo, and arithmetic operations are performed in parallel on each modulo [10]. The authors of [11] propose a method for constructing digital filters in RNS to automate the device design process and provide an effective speed and energy efficiency ratio. In [12], a new architecture for multiply-accumulate (MAC) units are proposed, which are the basis of digital filters. The proposed architecture is based on ternary value logic and RNS. However, using this approach leads to the high complexity of converting between ternary value logic and RNS. The authors of the paper [13] proposed a filter architecture with
finite impulse response based on truncated MAC units (TMAC). In [14], the implementation of TMAC units in RNS with moduli of a special type \(2^a\) and \(2^a - 1\), \(a \in \mathbb{N}\), where \(\mathbb{N}\) is the natural numbers set. The moduli of a special type allow reducing the operation of calculating the remainder of a division to a bit shift operation (for modulo \(2^a\)) and an addition operation of \(a\)-bit numbers (for modulo \(2^a - 1\)) [15], and use efficient addition and multiplication techniques [16, 17].

In this work, the device architecture for two-dimensional filtering by Winograd method for a filter mask \(2 \times 2\) using RNS with the moduli of the special type \(2^a\) and \(2^a - 1\). In the experimental part of the paper, a theoretical analysis of the proposed devices’ delay and area parameters is carried out. The theoretical results are confirmed by hardware simulation on FPGA.

The rest of the article is organized in the following way. In the second section, features of digital filtering in RNS are presented. In the third section, the Winograd method for two-dimensional signal filtering is described. A new device architecture for filtering by Winograd method using calculation in RNS by moduli \(2^a\) and \(2^a - 1\) is presented in the fourth section. The fifth section contains theoretical analysis results, hardware simulation on FPGA, and comparison with known digital filter architectures. The research results analysis and their discussion are carried out in the sixth section. The conclusions are presented in the seventh section.

II. DIGITAL FILTERING IN RESIDUE NUMBER SYSTEM

Digital filtering is applied for digital signal processing. In the case of processing a one-dimensional signal consisting of \(N\) samples by the filter \(f\) of size \(k\), filtering is described by the following formula [18]:

\[
y_l = \sum_{i=0}^{k-1} x_i \cdot f_{l-i},
\]

(1)

where \(y\) is a processing signal, \(0 \leq l < N\). When processing the two-dimensional signal \(X\) consisting of \(N \times K\) samples using \(k \times k\) filter \(F\), the filtering has the form

\[
y_{lt} = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} x_{ij} \cdot F_{l-i,j},
\]

(2)

where \(0 \leq l < N, 0 \leq t < K\).

As seen from (1) and (2), signal filtering contains addition and multiplication operations. The main computational load is multiply performing the multiplication operation. One of the ways to increase the performance of digital filtering devices is performing computations in RNS. In RNS, numbers are represented on the basis of coprime numbers, called moduli \(\{p_1, \ldots, p_n\}\), \(\text{GCD}(p_i, p_j) = 1\) for \(i \neq j\). All RNS moduli product \(P = \prod_{i=1}^{n} p_i\) is called the system dynamic range. Any integer \(0 \leq X < M\) it is uniquely represented in RNS as a vector \(\{x_1, x_2, \ldots, x_n\}\), where \(x_j = |X|_{p_j}\) is a remainder under division \(X\) by modulo \(p_j\) [10].

Fig. 1 shows the scheme of the signal \(x\) filtering in RNS, which consists of three stages. A first stage is a conversion number from the positional number system (PNS).

**FIGURE 1.** The scheme of filtering in RNS.

Then, filtering is performed in parallel for each modulo. Since addition, subtraction, and multiplication operation in RNS are determined by the formulas:

\[
A + B = \left( \left\lfloor a_1 + b_1 \right\rfloor_{p_1}, \ldots, \left\lfloor a_\eta + b_\eta \right\rfloor_{p_\eta} \right),
\]

\[
A \times B = \left( \left\lfloor a_1 \times b_1 \right\rfloor_{p_1}, \ldots, \left\lfloor a_\eta \times b_\eta \right\rfloor_{p_\eta} \right),
\]

(3)

then the one-dimensional filtration represented by expression (1) modulo \(p_i\) has the following form [18]:

\[
\left\lfloor y_i \right\rfloor_{p_i} = \left\lfloor \sum_{i=0}^{k-1} x_i \cdot f_{i-l} \right\rfloor_{p_i}.
\]

(4)

Similarly to (4), for the two-dimensional case (2), signal processing modulo \(p_i\) is described as follows:

\[
\left\lfloor y_{lt} \right\rfloor_{p_i} = \left\lfloor \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} x_{ij} \cdot F_{l-i,j} \right\rfloor_{p_i}.
\]

(5)

The last stage performs the reverse RNS to PNS transforms. Reconstruction of the number \(y\) from residuals \(\{y_1, y_2, \ldots, y_\eta\}\) is based on the Chinese remainder theorem [19]:

\[
y = \sum_{i=0}^{\eta} \left\lfloor p_i^{-1} \right\rfloor_{p_i} y_i \left\lfloor P_i \right\rfloor_{p_i},
\]

(6)

where \(P_i = P/p_i\). The term \(\left\lfloor P_i^{-1} \right\rfloor_{p_i}\) means multiplicative inverse for \(P_i\) by modulo \(p_i\).

The RNS moduli type affects the device’s technical characteristics, such as performance, hardware costs, and power consumption. Moduli of the special type \(2^a\) and \(2^a - 1\), \(a \in \mathbb{N}\), where \(\mathbb{N}\) is the natural numbers set, avoid division operation, which requires extensive computational resources [17].

In this paper, we propose the architecture of a two-dimensional filter with calculations by the Winograd method in RNS with moduli of the special type \(2^a\) and \(2^a - 1\). The operations of conversion to RNS and reverse transform to PNS were not considered in this study.
III. COMPUTATIONS IMPLEMENTATION IN DIGITAL FILTERS BY WINGRAD METHOD

One-dimensional filtering by Winograd method in matrix terms has the form:

\[ z = A^T((Gf) \odot (B^Td)), \]

(7)

where the operator \( \odot \) denotes elementwise matrix multiplication, \( A, G, \) and \( B \) are transformation matrices, \( f \) is a one-dimensional filter mask, \( d \) is a data vector, \( z \) is filtering result [8]. The one-dimensional filtering algorithm by Winograd method is usually denoted \( F(n, k) \) where \( n \) is a size of vector \( z \) and \( k \) is a filter \( f \) mask size.

Two-dimensional filtering by Winograd method in matrix terms has the form [8]:

\[ z = A^T((GfG^T) \odot (B^TdB))A, \]

(8)

where \( f, d \) and \( z \) are two-dimensional matrices. The two-dimensional filtering algorithm by Winograd method is usually denoted \( F(n \times n, k \times k) \).

Consider one-dimensional filtering by Winograd method using the example \( F(2, 2) \) [8]. Let us represent the vectors \( f, d, \) and \( z \) as polynomials:

\[
\begin{align*}
    f(x) &= f_1 x + f_0, \\
    z(x) &= z_1 x + z_0, \\
    d(x) &= d_2 x^2 + d_1 x + d_0.
\end{align*}
\]

(9)

Then the filtering is represented as a product of polynomials:

\[ d(x) = f(x)z(x). \]

(10)

We represent the polynomial \( d(x) \) as the remainder under division by the polynomial \( m(x) \) of the fourth degree

\[ d(x) = f(x)z(x) \mod m(x), \]

(11)

here \( \mod \) is a modulo polynomial division operator.

If we replace \( m(x) \) of the fourth degree by a polynomial of the third degree, then

\[ d(x) = f(x)z(x) \mod m(x) + R_{m(x)}[d(x)], \]

(12)

where \( R_{m(x)}[d(x)] \) is the remainder of \( d(x) \) divided by \( m(x) \).

We choose a polynomial \( m(x) = m^{(0)}(x)m^{(1)}(x)m^{(2)}(x) = x(x-1)(x-\infty) \), where \( x-\infty \) corresponds to \( R_{m(x)}[d(x)] \) [7, 8]. Then the remainders of division \( f(x) \) by \( m^{(i)}(x) \) are

\[
\begin{align*}
    f^{(0)}(x) &= f(x) \mod m^{(0)}(x) = f_0, \\
    f^{(1)}(x) &= f(x) \mod m^{(1)}(x) = f_0 + f_1, \\
    f^{(2)}(x) &= f(x) \mod m^{(2)}(x) = f_1.
\end{align*}
\]

(13)

And the remainders of division \( z(x) \) by \( m^{(i)}(x) \) are

\[
\begin{align*}
    z^{(0)}(x) &= z(x) \mod m^{(0)}(x) = z_0, \\
    z^{(1)}(x) &= z(x) \mod m^{(1)}(x) = z_0 + z_1, \\
    z^{(2)}(x) &= z(x) \mod m^{(2)}(x) = z_1.
\end{align*}
\]

(14)

The transformation matrix \( A \) is composed of the coefficients at the remainder of the division \( z(x) \) by \( m^{(i)}(x) \) and has the following form:

\[ A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}. \]

(15)

Let \( M^{(i)}(x) = \frac{m(x)}{m^{(i)}(x)} \), then

\[
\begin{align*}
    M^{(0)}(x) &= x - 1, \\
    M^{(1)}(x) &= x, \\
    M^{(2)}(x) &= x^2 - x.
\end{align*}
\]

(16)

The transformation matrix \( B \) consists of the coefficients of the polynomials \( M^{(i)}(x) \) and \( m(x) \), so that the coefficients \( M^{(i)}(x) \) correspond to the \( i \)-th column of the matrix \( B \)

\[ B = \begin{bmatrix} -1 & 0 & 0 \\ 1 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}. \]

(17)

According to CRT [7], it is necessary to choose such \( h^{(i)}(x) \) and \( H^{(i)}(x) \), that \( h^{(i)}(x)m^{(i)}(x) + H^{(i)}(x)M^{(i)}(x) = 1 \):

\[
\begin{align*}
    h^{(0)}(x) &= 1, & H^{(0)}(x) &= -1; \\
    h^{(1)}(x) &= -1, & H^{(1)}(x) &= 1.
\end{align*}
\]

(18)

The transformation matrix \( G \) composed of the reminders coefficients \( f^{(i)}(x) \) multiplied by \( H^{(i)}(x) \):

\[ G = \begin{bmatrix} -1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}. \]

(19)

For two-dimensional filtering \( F(2 \times 2, 2 \times 2) \) calculations are performed by formula (8). The following are the device architecture for two-dimensional filtering according to the Winograd method \( F(2 \times 2, 2 \times 2) \) with calculations in RNS.

IV. THE PROPOSED FILTER ARCHITECTURE WITH CALCULATIONS BY WINGRAD METHOD IN THE RESIDUE NUMBER SYSTEM

We divide the two-dimensional signal \( x \) into fragments \( d \) of size \( m \times m, m > k \). Each fragment is processed by the filter \( f \) of dimension \( k \times k \) according to Winograd method \( F(n \times n, k \times k) \) with a step \( n \) in each dimension.

\[ \text{FIGURE 2. The scheme of filtering a fragment of the two-dimensional signal by the Winograd method.} \]

In the case of \( F(2 \times 2, 2 \times 2) \), the two-dimensional signal \( x \) is divided into \( 3 \times 3 \) fragments, and the processing is...
performed with step 2. Fig. 2 shows a filtering process scheme for the fragment $d$ according to the Winograd method $F(2 \times 2, 2 \times 2)$.

The two-dimensional filtering by the Winograd method described by (8) performs signal processing in several stages. The filter mask transformation result is denoted as $U = GfG^T$. Since the filter coefficients are constants, this transformation is performed in advance and does not carry the computational load. The input data $d$ transformation result is denoted as $V = B^TdB$, and the element-wise matrix multiplication result is $M = U \odot V$. Then, taking into account the introduced notation, formula (8) takes the form $z = A^TMA$.

When filtering by Winograd method $F(2 \times 2, 2 \times 2)$, the matrix $V$ and the final transformation matrix $M$ are formed by the following way:

$$
V = \begin{bmatrix}
    d_{0,0} - d_{1,0} - d_{0,1} + d_{1,1} & d_{1,0} - d_{1,1} - d_{0,2} + d_{1,2} \\
    -d_{1,0} + d_{0,1} & -d_{0,1} + d_{1,2} \\
    d_{1,0} - d_{2,0} - d_{1,1} + d_{2,1} & d_{1,1} - d_{2,1} - d_{1,2} + d_{2,2}
\end{bmatrix},
$$

(20)

$$
z = \begin{bmatrix}
    M_{0,0} + M_{1,0} + M_{0,1} + M_{1,1} \\
    M_{1,0} + M_{2,0} + M_{1,1} + M_{2,1} \\
    M_{1,1} + M_{1,2} + M_{2,1} + M_{2,2}
\end{bmatrix}.
$$

(21)

The modulo $2^a$ and $2^a - 1$ addition of several numbers is performed using a multiply modulo adder (MOMA), we introduce for them the notation $MOMA_{2^a}$ and $MOMA_{2^a-1}$, respectively (Fig. 3). These devices consist of a carry-save adder (CSA) [20] and Kogge-Stone Adder (KSA) [21]. The vector $P = \{P_0, P_1, ..., P_p\}$ is the input of the devices, and the sum $S$ is the output. An end-around-carry (EAC) technique is used for modulo $2^a - 1$ calculations [17].

**FIGURE 3.** The scheme of MOMA device: a) modulo $2^a$ ($MOMA_{2^a}$); b) modulo $2^a - 1$ ($MOMA_{2^a-1}$).

The matrix $V$ elements modulo $2^a$ are calculated using the device $MOMA_{2^a}$ (Fig. 3a). The representation of negative numbers modulo $2^a$ requires converting them into two’s complement code, that is inverting the number and adding one. Therefore, the input $MOMA_{2^a}$ is fed with the data vector $D^{i,j}$ and the correction constant $C^{i,j}$, which is equal to the negative numbers amount, to calculate the element $V_{i,j}$ modulo $2^a$ where $0 \leq i \leq 2$ and $0 \leq j \leq 2$. An exception is an element $V_{1,1} = D^{1,1} = d_{0,1}$, which does not require any calculations. Thus, the data transformation device modulo $2^a$ (denote it $DT_{2^a}$) consists of eight devices $MOMA_{2^a}$, the inputs of which receive the following data:

$$
D^{0,0} = \{d_{0,0}, \overline{d_{1,0}}, \overline{d_{0,1}}, d_{1,1}\}, C^{0,0} = 2;
$$

$$
D^{0,1} = \{d_{0,0}, d_{1,1}\}, C^{0,1} = 1;
$$

$$
D^{0,2} = \{d_{1,0}, d_{1,1}, d_{0,2}, d_{1,2}\}, C^{0,2} = 2;
$$

$$
D^{1,0} = \{\overline{d_{1,0}}, d_{0,1}\}, C^{1,0} = 1;
$$

$$
D^{1,1} = \{\overline{d_{1,0}}, \overline{d_{1,1}}, d_{2,1}\}, C^{1,1} = 2;
$$

$$
D^{2,0} = \{d_{1,0}, d_{2,0}, d_{1,1}, d_{2,1}\}, C^{2,0} = 2;
$$

$$
D^{2,1} = \{d_{1,1}, d_{2,1}\}, C^{2,1} = 1;
$$

$$
D^{2,2} = \{d_{1,1}, \overline{d_{2,1}}, \overline{d_{1,2}}, d_{2,2}\}, C^{2,2} = 2.
$$

(22)

The matrix $V$ elements modulo $2^a - 1$ calculation requires the representation of negative numbers in the one’s complement code, which is the inversion of the number. Therefore, the correction constants are not involved in the calculations. As seen from (22), the data transformation device modulo $2^a - 1$ (denote it $DT_{2^a-1}$) consists of four devices...
MOMA\(_{2\alpha-1}\), the inputs of which receive vectors \(D^{0,0}\), \(D^{0,2}\), \(D^{2,0}\) and \(D^{2,2}\), and of four EAC-KSA, to whose inputs are fed with vectors \(D^{0,1}\), \(D^{1,0}\), \(D^{1,2}\) and \(D^{2,1}\).

The elementwise matrices \(U\) and \(V\) multiplication devices modulo \(2^\alpha\) and \(2^\alpha - 1\) consist of nine multipliers \(MUL_{2\alpha}\) and \(MUL_{2\alpha-1}\), respectively, shown in Fig. 4. The multiplier \(MUL_{2\alpha}\) consists of a partial product generator modulo \(2^\alpha\) \(PPG_{2\alpha}\), which is formed from an array of AND gates [20], and \(MOMA_{2\alpha}\). The \(MUL_{2\alpha-1}\) device consists of the modulo \(2^\alpha - 1\) partial product generator \(PPG_{2\alpha-1}\) using the EAC technique, and \(MOMA_{2\alpha-1}\). We denote the elementwise matrices multiplication device modulo \(2^\alpha\) as \(EWM_{2\alpha}\), and modulo \(2^\alpha - 1\) as \(EWM_{2\alpha-1}\). Matrix \(M\) is formed at the outputs of these devices.

**FIGURE 4.** The multiplier scheme: a) modulo \(2^\alpha\) (\(MUL_{2\alpha}\)); b) modulo \(2^\alpha - 1\) (\(MUL_{2\alpha-1}\)).

The matrix \(z\) is formed by final transformation devices modulo \(2^\alpha\) and \(2^\alpha - 1\), which denoted \(FT_{2\alpha}\) and \(FT_{2\alpha-1}\). The \(FT_{2\alpha}\) device consists of four \(MOMA_{2\alpha}\), their inputs are vectors \(R_{ij}^{\alpha}\):

\[
\begin{align*}
R_{ij}^{0,0} &= \{M_{0,0}, M_{1,0}, M_{0,1}, M_{1,1}\}, \\
R_{ij}^{0,1} &= \{M_{0,1}, M_{1,1}, M_{0,2}, M_{1,2}\}, \\
R_{ij}^{1,0} &= \{M_{1,0}, M_{2,0}, M_{1,1}, M_{2,1}\}, \\
R_{ij}^{1,1} &= \{M_{1,1}, M_{2,1}, M_{1,2}, M_{2,2}\},
\end{align*}
\]

and the outputs are \(z_{ij}\), where \(0 \leq i \leq 2\) and \(0 \leq j \leq 2\). The \(FT_{2\alpha-1}\) device consists of four \(MOMA_{2\alpha-1}\) the inputs of which also receive vectors \(R_{ij}^{\alpha}\).

Therefore, the \(F(2 \times 2 \times 2)_{2\alpha}\) device for filtering by Winograd method modulo \(2^\alpha - 1\) has a similar structure and consists of \(DT_{2\alpha-1}\), \(EWM_{2\alpha-1}\) and \(FT_{2\alpha-1}\) units (Fig. 6). The data fragment \(d_{ij}^{\alpha-1}\) enters at the device input, and the processed fragment \(z_{ij}^{\alpha-1}\) is formed at the output.

Fig. 7 presents the proposed \(F(2 \times 2 \times 2)_{\text{RNS}}\) device architecture for filtering by Winograd method using calculations in RNS with moduli set \(\{2^\alpha, 2^\alpha - 1, \..., 2^\alpha \eta - 1\}\). The data array \(d_{\text{RNS}} = \{d_{1}^{\alpha-1}, \..., d_{2\alpha-1}^{\alpha-1}\}\) presented in RNS enters at the device input. Then, the input data is processed in parallel by \(F(2 \times 2 \times 2)_{2\alpha}\). \(F(2 \times 2 \times 2)_{2\alpha-1}\), \(F(2 \times 2 \times 2)_{2\alpha-1}\) devices which are formed the processed data array \(z_{\text{RNS}} = \{z_{1}^{\alpha}, \..., z_{2\alpha-1}^{\alpha-1}\}\) at the output. The main difference between the proposed filtering method based on the Winograd method and the method presented in [9] is the execution of computations in the RNS with the special type moduli set.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
M. Valueva et al.: Digital Filter Architecture with Calculations in the RNS by Winograd Method $F(2 \times 2 \times 2)$

**FIGURE 5.** The $F(2 \times 2 \times 2)$ device for two-dimensional filtering by the Winograd method modulo $2^m$.

**FIGURE 6.** The $F(2 \times 2 \times 2)_{2^m-1}$ device for two-dimensional filtering by the Winograd method modulo $2^m - 1$. 
In the next section, we perform the parameters theoretical analysis of the proposed device for filtering by the Winograd method using RNS arithmetic.

V. RESEARCH RESULTS

A. THEORETICAL ANALYSIS

We use the abstract "unit-gate" model to estimate digital devices' delay and area parameters [22]. According to this model, if we denote the logic device delay $U_{\text{delay}}$ and denote the logical device area as $U_{\text{area}}$, then the logic gates descriptions are

$$
U_{\text{delay}}(\text{NOT}) = 0, U_{\text{area}}(\text{NOT}) = 0; \\
U_{\text{delay}}(\text{AND}) = 1, U_{\text{area}}(\text{AND}) = 1; \\
U_{\text{delay}}(\text{OR}) = 1, U_{\text{area}}(\text{OR}) = 1; \\
U_{\text{delay}}(\text{XOR}) = 2, U_{\text{area}}(\text{XOR}) = 2; \\
U_{\text{delay}}(\text{XNOR}) = 2, U_{\text{area}}(\text{XNOR}) = 2.
$$

Then, according to (24), $MOMA_{2^\alpha}$ parameters have the form, where $N$ is the number of terms [14]:

$$
U_{\text{delay}}(MOMA_{2^\alpha}) = 6.8 \log_2 N + 2 \log_2 2 + 4, \\
U_{\text{area}}(MOMA_{2^\alpha}) = 3 a \log_2 2 + 7 a N - 11 a + 1.
$$

Similarly, $MOMA_{2^{\alpha-1}}$ parameters are represented as [14]:

$$
U_{\text{delay}}(MOMA_{2^{\alpha-1}}) = 6.8 \log_2 N + 2 \log_2 2 + 4, \\
U_{\text{area}}(MOMA_{2^{\alpha-1}}) = 3 a \log_2 2 + 7 a N - 8 a.
$$

Data conversion device $DT_{2^\alpha}$ contains four $MOMA_{2^\alpha}$, at the input of which five terms are received, and four $MOMA_{2^\alpha}$, at the input of which three terms are received. Then, taking into account (25), $DT_{2^\alpha}$ device parameters are calculated as follows:

$$
U_{\text{delay}}(DT_{2^\alpha}) = 2 \log_2 2 + 24.4, \\
U_{\text{area}}(DT_{2^\alpha}) = 24 a \log_2 2 + 136 a + 8.
$$

Data conversion device $DT_{2^{\alpha-1}}$ consists of four $MOMA_{2^{\alpha-1}}$, at the input of which four terms are received, and four EAC-KSA. Then, based on (24) and (26), parameters of the $DT_{2^{\alpha-1}}$ device are:

$$
U_{\text{delay}}(DT_{2^{\alpha-1}}) = 2 \log_2 2 + 24.4, \\
U_{\text{area}}(DT_{2^{\alpha-1}}) = 24 a \log_2 2 + 16 a.
$$

The multiplier consists of a partial product generator and MOMA unit for $\alpha$ inputs. Partial product generator $PPG_{2^\alpha}$ has the following delay and area parameters:

$$
U_{\text{delay}}(PPG_{2^\alpha}) = 0.5 a^2 + 0.5 a, \\
U_{\text{area}}(PPG_{2^\alpha}) = 0.5 a^2 + 0.5 a.
$$

Partial product generator $PPG_{2^{\alpha-1}}$ has the following parameters:

$$
U_{\text{delay}}(PPG_{2^{\alpha-1}}) = a^2, \\
U_{\text{area}}(PPG_{2^{\alpha-1}}) = a^2.
$$

Then, taking into account (25) and (29), $MUL_{2^\alpha}$ has the following parameters:

$$
U_{\text{delay}}(MUL_{2^\alpha}) = 8.8 \log_2 2 + 0.5 a^2 + 0.5 a + 4, \\
U_{\text{area}}(MUL_{2^\alpha}) = 3a \log_2 2 + 7.5 a^2 - 10.5 a + 1.
$$

And $MUL_{2^{\alpha-1}}$, based on (26) and (30), has parameters:

$$
U_{\text{delay}}(MUL_{2^{\alpha-1}}) = 8.8 \log_2 2 + a^2 + 4, \\
U_{\text{area}}(MUL_{2^{\alpha-1}}) = 3a \log_2 2 + 8 a^2 - 8 a.
$$

The elementwise matrices multiplication device $EW M_{2^\alpha}$ consists of nine $MUL_{2^\alpha}$ with parameters (31). The $EW M_{2^\alpha}$ device has the following parameters:

$$
U_{\text{delay}}(EW M_{2^\alpha}) = 8.8 \log_2 2 + 0.5 a^2 + 0.5 a + 4, \\
U_{\text{area}}(EW M_{2^\alpha}) = 27 a \log_2 2 + 67.5 a^2 - 90.5 a + 9.
$$

$EW M_{2^{\alpha-1}}$ device consists of nine $MUL_{2^{\alpha-1}}$ with parameters (32). Therefore, its parameters are represented as:

$$
U_{\text{delay}}(EW M_{2^{\alpha-1}}) = 8.8 \log_2 2 + a^2 + 4, \\
U_{\text{area}}(EW M_{2^{\alpha-1}}) = 27 a \log_2 2 + 72 a^2 - 72 a.
$$

The final transformation $FT_{2^\alpha}$ device consists of four $MOMA_{2^\alpha}$ devices, at the input of which four terms are received. Then, taking into account $MOMA_{2^\alpha}$ parameters (25), $FT_{2^\alpha}$ device has the following parameters:

$$
U_{\text{delay}}(FT_{2^\alpha}) = 2 \log_2 2 + 17.6, \\
U_{\text{area}}(FT_{2^\alpha}) = 12 a \log_2 2 + 68 a^2 + 4.
$$

Similarly, $FT_{2^{\alpha-1}}$ device consists of four $MOMA_{2^{\alpha-1}}$ devices, at the input of which four terms are received. Considering $MOMA_{2^{\alpha-1}}$ parameters (24), $FT_{2^{\alpha-1}}$ device has the following parameters:

$$
U_{\text{delay}}(FT_{2^{\alpha-1}}) = 6.8 \log_2 2 + 2 \log_2 2 + 4, \\
U_{\text{area}}(FT_{2^{\alpha-1}}) = 12 a \log_2 2 + 80 a.
$$

Taking into account parameters of the $DT_{2^\alpha}$, $EW M_{2^\alpha}$ and $FT_{2^\alpha}$ devices – (27), (33), (36), the filter $F(2 \times 2 \times 2)_{2^\alpha}$ device based on Winograd method with modulo $2^\alpha$ calculations has the following delay and area parameters:
The delay and area theoretical parameters of the filter based on Winograd method with modulo $2^n-1$ calculation has the following delay and area parameters:

\[
\begin{align*}
U_{\text{delay}}(F(2 \times 2,2 \times 2)) &= 12.8 \log_2 \alpha + 0.5 \alpha^2 + 0.5 \alpha + 46, \\
U_{\text{area}}(F(2 \times 2,2 \times 2)) &= 63 \alpha \log_2 \alpha + 67.5 \alpha^2 - 113.5 \alpha + 21.
\end{align*}
\]

Parameters of the proposed filtering device $F(2 \times 2,2 \times 2)$ with calculations in the RNS with moduli set $\{2^a_1, 2^a_2 - 1, ..., 2^a_n - 1\}$, shown in Fig. 6, taking into account (37) and (38), are calculated as follows:

\[
\begin{align*}
U_{\text{delay}}(F(2 \times 2,2 \times 2)) &= \max \{U_{\text{delay}}(F(2 \times 2,2 \times 2)_a), \\
&\quad \{U_{\text{delay}}(F(2 \times 2,2 \times 2)_{a_i})\} | i \leq \eta\}, \\
U_{\text{area}}(F(2 \times 2,2 \times 2)) &= \sum_{i=2}^n U_{\text{area}}(F(2 \times 2,2 \times 2)_{a_i} + U_{\text{area}}(F(2 \times 2,2 \times 2)_{a_i-1}).
\end{align*}
\]

Parameters of the proposed filtering device $F(2 \times 2,2 \times 2)$ with calculations in the RNS with moduli set $\{2^a_1, 2^a_2 - 1, ..., 2^a_n - 1\}$, shown in Fig. 6, taking into account (37) and (38), are calculated as follows:

\[
\begin{align*}
U_{\text{delay}}(F(2 \times 2,2 \times 2)_{RNS}) &= \max \{U_{\text{delay}}(F(2 \times 2,2 \times 2), \\
&\quad \{U_{\text{delay}}(F(2 \times 2,2 \times 2)_{a_i})\} | i \leq \eta\}, \\
U_{\text{area}}(F(2 \times 2,2 \times 2)) &= \sum_{i=2}^n U_{\text{area}}(F(2 \times 2,2 \times 2)_{a_i} + U_{\text{area}}(F(2 \times 2,2 \times 2)_{a_i-1}).
\end{align*}
\]

A theoretical analysis of the proposed device delay and area parameters based on the "unit-gate" model using various RNS moduli sets is performed, as well as a comparison with known filter architectures. Also, the processing time for a $256 \times 256$
two-dimensional signal fragment is calculated according to the device delay. The filter parameters calculations results are presented in Table II.

Delay and area parameters theoretical analysis based on the "unit-gate" model of the proposed filter device and known analogs allows to conclude about a computational \(O_{\text{delay}}\) and space \(O_{\text{area}}\) complexity of reviewed methods. Let the bit width of the RNS dynamic range be approximately \(\alpha\), and the bit width of each computational channel is approximately \(\frac{\alpha}{\eta}\).

Then, according to (39), the proposed filtering algorithm complexity by the Winograd method in RNS is

\[
O_{\text{delay}}(F(2 \times 2 \times 2)_{\text{RNS}}) = \left(\frac{\alpha}{\eta}\right)^2, \quad O_{\text{area}}(F(2 \times 2 \times 2)_{\text{RNS}}) = \left(67.5 + 72(\eta - 1)\right)\left(\frac{\alpha}{\eta}\right)^2. \tag{44}
\]

The complexity of the filtering algorithm by the Winograd method [9], according to (37), is

\[
O_{\text{delay}}(F(2 \times 2 \times 2)_{2^2}) = 0.5\alpha^2, \quad O_{\text{area}}(F(2 \times 2 \times 2)_{2^2}) = 67.5\alpha^2. \tag{45}
\]

Comparing the complexity of the proposed method (44) and the method [9] (45), we can conclude that if the number of modules from two or more \((\eta \geq 2)\), then the computational and space complexity of the proposed method is less than the method [9], that is \(O_{\text{delay}}(F(2 \times 2 \times 2)_{\text{RNS}}) < O_{\text{delay}}(F(2 \times 2 \times 2)_{2^2})\) and \(O_{\text{area}}(F(2 \times 2 \times 2)_{\text{RNS}}) < O_{\text{area}}(F(2 \times 2 \times 2)_{2^2})\).

Since the output of the filter based on the Winograd method \(F(2 \times 2 \times 2)\) is four processed signal values, then for filters based on MAC [23] and TMAC [13,14] units, we calculate the complexity of the method from the calculation of sequential processing four signal fragments.

According to the "unit-gate" model parameters (40), the complexity of the filtering method based on MAC units [23] is

\[
O_{\text{delay}}(\text{FIR(MAC)}) = 176 \log_2 \alpha, \quad O_{\text{area}}(\text{FIR(MAC)}) = 40\alpha^2. \tag{46}
\]

Let us compare expressions (44) and (46). If \(\frac{\alpha^2}{144\eta^2\log_2 \alpha} < 1\), then the computational complexity of the proposed method is less than the method based on MAC units [23], that is \(O_{\text{delay}}(F(2 \times 2 \times 2)_{\text{RNS}}) < O_{\text{delay}}(\text{FIR(MAC)})\). If \(\eta \geq 2\), the space complexity of the proposed method is less than method [23], that is \(O_{\text{area}}(F(2 \times 2 \times 2)_{\text{RNS}}) < O_{\text{area}}(\text{FIR(MAC)})\).

According to delay and area parameters (41), the complexity of the filtering method based on TMAC units [13] is

\[
O_{\text{delay}}(\text{FIR(TMAC)}_{2^\alpha}) = 144 \log_2 \alpha, \quad O_{\text{area}}(\text{FIR(TMAC)}_{2^\alpha}) = 40\alpha^2. \tag{47}
\]

Let us compare expressions (44) and (47). If \(\frac{\alpha^2}{144\eta^2\log_2 \alpha} < 1\), then the computational complexity of the proposed method is less than the method based on TMAC units [13], that is \(O_{\text{delay}}(F(2 \times 2 \times 2)_{\text{RNS}}) < O_{\text{delay}}(\text{FIR(TMAC)}_{2^\alpha})\). If \(\eta \geq 2\) then the space complexity of the proposed method is less, that is \(O_{\text{area}}(F(2 \times 2 \times 2)_{\text{RNS}}) < O_{\text{area}}(\text{FIR(TMAC)}_{2^\alpha})\).

According to the "unit-gate" model parameters (43), the complexity of the TMAC-based method with calculations in RNS [14] is

\[
O_{\text{delay}}(\text{FIR(TMAC)}_{\text{RNS}}) = 144 \log_2 \frac{\alpha}{\eta}, \quad O_{\text{area}}(\text{FIR(TMAC)}_{\text{RNS}}) = 40\frac{\alpha^2}{\eta}. \tag{48}
\]

Let us compare expressions (44) and (48). If \(\frac{\alpha^2}{144\eta^2\log_2 \alpha} < 1\), then the computational complexity of the proposed method is less than the method [14], that is \(O_{\text{delay}}(F(2 \times 2 \times 2)_{\text{RNS}}) < O_{\text{delay}}(\text{FIR(TMAC)}_{\text{RNS}})\). If any integer \(\eta > 1\), then the space complexity of the proposed method is greater than the method [14], that is \(O_{\text{area}}(F(2 \times 2 \times 2)_{\text{RNS}}) > O_{\text{area}}(\text{FIR(TMAC)}_{\text{RNS}})\).

A description of the theoretical analysis results is presented in VI Section "Discussion".

\section*{B. HARDWARE IMPLEMENTATION}

Hardware simulation of the proposed device architecture for filtering by the Winograd method \(F(2 \times 2 \times 2)\) with calculations in RNS with the special moduli type \(2^a\) and \(2^a - 1\). The proposed architecture was compared with known developments. The hardware simulation results are presented in Table III. The simulation was carried out in Xilinx Vivado 2018.3 environment for the target board Artix-7 xc7a200tfg1156-3 with optimization strategy Flow_Performimized_high. The following parameters were used to evaluate the devices: clock frequency, number of busy Look-Up-Tables (LUTs), power consumption, and performance equal to the number of processed fragments \(256 \times 256\) of two-dimensional signals per second (fragments/s). A description of the hardware implementation results is presented in VI Section "Discussion".

\section*{VI. DISCUSSION}

Comparison of the computational and space complexity of the proposed and known methods showed that with an increase in the dynamic range and with a decrease in the number of RNS modules, the advantage of the proposed method decreases. Since, in practice, signal processing systems with small bit width (for example, 8-, 16-, 32-bit) are more often used, these limitations are insignificant.
Theoretical analysis results (Table II) based on the "unit-gate" model of the proposed device parameters showed that RNS usage allows to reduce the device delay by 24.79% – 66.77%, and the area device by 17.59% – 53.67%, compared with the known implementation based on Winograd filtering in PNS [9]. In addition, the proposed device architecture has 13.47% – 42.04% less delay, and 2.20% – 18.03% less area, except for the 8-bit device, which has a 47.38% larger area than the known MAC-based filter architecture [23]. Compared to the known device architecture based on TMAC units with computations in PNS [13], the delay of the proposed device is 20.92% – 22.22% less, but the area is 1.56% – 53.37% more for 8- and 16-bit devices, and for 32-bit devices, the delay is 12.17% larger, but the area is 18.03% less. The proposed architecture of the 8-bit filter has 2.15% lower latency, but 16- and 32-bit devices have 3.42% – 52.52% more delay, compared to the known architecture based on TMAC units with computations in RNS. The area of the proposed device is approximate twice the area of a device based on TMAC units with computations in RNS [14]. The main advantage of the proposed filter architecture based on the Winograd method with calculations in RNS with special type modules is to reduce the processing time of a two-dimensional signal. Thus, the use of the proposed device makes it possible to reduce the processing time of a 256 × 256 signal by 1.33 – 6.90 times compared to other known architectures.

The results of hardware simulation (Table III) showed that the performance of the proposed filter architecture is 1.31 – 4.12 times higher in comparison with known architectures. The maximum clock frequency of the proposed device is 31.03% – 38.46% higher compared with device based on the Winograd method [9] and 1.89% – 2.94% higher compared to the filter based on MAC units [23] for the case of 8- and 16-bit devices, but 26.21% less for the case of 8-bit devices. Nevertheless, compared with the TMAC-based filter architecture in PNS [13] and RNS [14], the performance of the proposed device is 7.89% – 35.59% lower. The number of LUTs occupied by the proposed device is 18.08% – 37.27% less than the filter based on the Winograd method [9], but 3.83 – 7.74 times more than other reviewed known architectures. The power consumption of the proposed filtering device is 2.57% – 42.46% higher than the known devices. The insignificant difference between theoretical analysis results and hardware simulation results is due to the peculiarities of the "unit-gate" model, which does not consider fan-out devices.

The high performance of filters with Winograd method calculations is explained because the result is several processed elements at once (Fig. 2). However, performance gains come at the expense of increased hardware costs, such as the number of occupied LUTs and power consumption. Thus, the proposed architecture can be successfully applied in digital signal processing systems, in which performance is a crucial criterion. However, from the occupied area's point of view, an architecture based on TMAC units with calculations in RNS is preferable [14]. The MAC-based filter [23] is advisable to use in systems with low power consumption.

The proposed filter architecture can be applied as part of the filter with a larger mask and decimation [9]. For example, filter with kernel size 3 × 3 and decimation step two may be presented as two filters based on Winograd method $F(2 \times 2, 2, 2) \times (2, 2, 1)$ and $F(2, 2, 1, 1)$. An interesting direction of future research is the described approaches application to the implementation of filters for the filter based on Winograd method for other dimensions of filter masks $F(n \times n, k \times k)$ and their application in convolutional layers of convolutional neural networks.

### VII. CONCLUSION

In this paper, we proposed the device architecture for two-dimensional filtering by Winograd method $F(2 \times 2, 2)$ using calculations in RNS with moduli of the special type $2^\alpha$ and $2^\beta − 1$. The theoretical analysis is performed based on the "unit-gate"-model and shows that the speed of signal processing by the proposed device is 1.33 – 6.90 times higher than other known devices. Also, the hardware implementation on FPGA is performed, showing that the proposed device performance is 1.31 – 4.12 times higher than other known methods. The research results may be applied to increase the digital signal processing device technical characteristics and in the intellectual analysis systems for data preprocessing.

### ACKNOWLEDGMENT
The authors would like to thank the North Caucasus Federal University for supporting in the contest of projects competition of scientific groups and individual scientists of the North Caucasus Federal University.

REFERENCES

[1] N. T. Bui et al., “Real-Time Filtering and ECG Signal Processing Based on Dual-Core Digital Signal Controller System,” in IEEE Sensors Journal, vol. 20, no. 12, pp. 6492-6503, 15 June15, 2020, doi: 10.1109/JSEN.2020.2975006.

[2] Z. Deng, P. Xu, L. Xie, K.-S. Choi and S. Wang, “Transductive Joint-Knowledge-Transfer TSK FS for Recognition of Epileptic EEG Signals,” in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 26, no. 8, pp. 1481-1494, Aug. 2018, doi: 10.1109/TNSRE.2018.2850308.

[3] K. Kamirul, W. Hasbi, P. R. Hakim and A. H. Syafrudin, “Automatic Ship Recognition Chain on Satellite Multispectral Imagery,” in IEEE Access, vol. 8, pp. 221918-221931, 2020, doi: 10.1109/ACCESS.2020.3042702.

[4] G. Wang, B. Li, Y. Zhang and J. Yang, “Background Modeling and Referencing for Moving Cameras-Captured Surveillance Video Coding in HEVC,” in IEEE Transactions on Multimedia, vol. 20, no. 11, pp. 2921-2934, Nov. 2018, doi: 10.1109/TMM.2018.2829163.

[5] L. M. B. Alonzo, F. B. Chiasson, H. S. Co, N. T. Bugtai and R. G. Baldovino, “A Machine Learning Approach for Coconut Sugar Quality Assessment and Prediction,” 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), 2018, pp. 1-4, doi: 10.1109/HNICEM.2018.8666315.

[6] J. Lu et al., “Real-Time FPGA-Based Digital Signal Processing and Correction for a Small Animal PET,” in IEEE Transactions on Nuclear Science, vol. 66, no. 7, pp. 1287-1295, July 2019, doi: 10.1109/TNS.2019.2908220.

[7] S. Winograd, “Arithmetic complexity of computations,” Siam; 1980:33. DOI: 10.1137/1.9781611970364.

[8] A. Lavin and S. Gray, “Fast Algorithms for Convolutional Neural Networks,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4013-4021, doi: 10.1109/CVPR.2016.435.

[9] J. Yepez and S. Ko, “Stride 2 1-D, 2-D, and 3-D Winograd for Convolutional Neural Networks,” in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 28, no. 4, pp. 853-863, April 2020, doi: 10.1109/TVLSI.2019.2961602.

[10] A. Omondi, B. Premkumar, Residue Number Systems: Theory and Implementation. London, U.K.: Imperial College Press, 2007, p. 296

[11] G. C. Cardarilli, L. Di Nunzio, R. Fazzolari, A. Nannarelli, M. Petricca and M. Re, "Design Space Exploration based Methodology for Residue Number System Digital Filters Implementation," in IEEE Transactions on Emerging Topics in Computing, doi: 10.1109/TETC.2020.2997067.

[12] A. Ghosh and A. Sinha, "FPGA Implementation of RNS Adder Based MAC Unit in Ternary Value Logic Domain for Signal Processing Algorithm and its Performance Analysis," 2018 IEEE Electron Devices Kozlata Conference (EDKCON), 2018, pp. 182-187, doi: 10.1109/EDKCON.2018.8770463.

[13] P. Lyakhov, M. Valueva, G. Valuev, N. Nagornov, "A Method of Increasing Digital Filter Performance Based on Truncated Multiply-Accumulate Units," Applied Sciences 2020, vol. 10, 9052, doi: /10.3390/app102409052.

[14] P. Lyakhov, M. Valueva, G. Valuev and N. Nagornov, “High-Performance Digital Filtering on Truncated Multiply-Accumulate Units in the Residue Number System,” in IEEE Access, vol. 8, pp. 209181-209190, 2020, doi: 10.1109/ACCESS.2020.3038496.

[15] P. M. Matutino, R. Chaves and L. Sousa, “Arithmetic-Based Binary-to-RNS Converter Module $S(\{2^n[a]\})S$ for $\text{SNR} - \text{bit Dynamic Range},$ in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 23, no. 3, pp. 603-607, March 2015, doi: 10.1109/TVLSI.2014.2314174.