Int. J. Electron. Commun. (AEÜ) 74 (2017) 13–20

Contents lists available at ScienceDirect

International Journal of Electronics and Communications (AEÜ)

journal homepage: www.elsevier.com/locate/aeue

Regular paper

Double span floating point representation scheme for the powers-of-two filter

Abhijit Chandra [*], Priya Kumari, Mouban Chakraborty, Utsab Sengupta

Department of Instrumentation and Electronics Engineering, Jadavpur University, Kolkata 700 098, India

ARTICLE INFO

Article history:
Received 12 September 2016
Accepted 24 January 2017

Keywords:
Double span floating point (DSFP)
Full adder count
Minimum index floating point (MIFP)
Powers-of-two filter
Pseudo floating point (PFP)

ABSTRACT

The area of signal processing has been experiencing silent revolution over the last few years. A number of promising algorithms are being developed in this regard. In connection to this, minimization of hardware complexity of digital filter has grown sufficient interest amongst the research community. Hardware cost of digital filter may be reduced by encoding the filter coefficient in the form of sum of signed powers-of-two (SPT). This article introduces a new encoding strategy of the non-uniform powers-of-two coefficients for the sake of exploiting minimum hardware units. Proposed scheme targets to minimize the highest powers-of-two terms in any coefficient by judiciously dividing the 'span' part into two segments. As a matter of fact, it necessitates the use of minimum number of full-adder blocks during implementation as compared to other existing coefficient representation schemes. Supremacy of the proposed double span floating point (DSFP) representation technique has been mathematically substantiated and supported with the help of few design examples.

1. Introduction

Finite impulse response (FIR) filter is gaining paramount importance in mobile communication systems to perform various functionalities such as channelization, channel equalization, matched filtering and pulse shaping due to its absolute stability and linear phase property [1]. It has also become an important tool for Digital Signal Processing (DSP) unit which has developed a number of portable battery powered wireless devices like mobile phones, laptops, wireless modems and so on for which the primary concern is system complexity and delay. Although programmable filters may surmount these problems, they are not suitable for their low throughput rate and high power consumption. As a matter of fact, dedicated FIR filters with small area and low power consumption has received justified attention over the last few decades. Since, the multipliers in the FIR architecture consume enormous power and acquire excessive area; they are substituted by shifters and adders and thus results in multiplier-less FIR filters. Complexity of any multiplier-less FIR filter is essentially determined by the number of full adders for its implementation as the shift elements are less complex and can be easily hardwired.

Design of multiplier-less filter has recently been addressed by several researchers throughout the globe for the purpose of minimizing the power consumption [2]. Different techniques have already been proposed in literature for the efficient realization of low-power FIR filter with minimum number of adders [3,4]. Several coefficient representation schemes like canonic signed digit (CSD), signed powers-of-two (SPT) [5,6] had been introduced in this regard. Hardware friendly fast parallel FIR filters are very recently designed by several means like cyclotomic polynomial [7], computer algebra system [8], iterated short convolution [9] and so on.

Common subexpression elimination (CSE) technique [10] emerges to be one of the most attractive representation strategies which multiply one variable (input signal) with several constants (filter coefficients). CSE technique uses the most commonly occurring subexpressions/bit patterns which exist in the CSD representation of coefficients and thus eliminates redundant computations in multiplier blocks. However, it is extremely expensive to pipeline due to its irregular nature of the subexpression elimination tree. Overall filters produced using the CSE technique therefore consumes more power since it uses the input signals and coefficients directly.

As an alternative to CSE technique, the differential coefficients method (DCM) was introduced in [11]. Instead of multiplying the coefficients directly, DCM multiplies differential coefficients with

[*] Corresponding author at: Department of Instrumentation and Electronics Engineering, Jadavpur University, Sector-III, Block LB, Plot No. 8, Salt Lake City, Salt Lake Bypass, Kolkata 700 098, India.

E-mail addresses: abhijit922@yahoo.co.in (A. Chandra), k.priya093@gmail.com (P. Kumari), moubanchakraborty@gmail.com (M. Chakraborty), utsab10manu@gmail.com (U. Sengupta).

http://dx.doi.org/10.1016/j.aeue.2017.01.018
1434-8411/© 2017 Elsevier GmbH. All rights reserved.

inputs. Since the difference in consecutive coefficients has shorter word length than the actual coefficient, DCM proves to be effective in reducing the power consumption. However, DCM requires extra adders to compute the sum of the stored partial products in order to compensate the effect of differential coefficients, and hence suffers from overhead. A new algorithm, called the differential coefficients and input method (DCIM), has been proposed in [12] to reduce the power consumption of filters which considers differential inputs as well as differential coefficients for its implementation. DCIM also suffers from the same overhead problem as that of DCM and it causes propagation delay after input arrival in order to derive the differential input. The minimal difference differential coefficients method (MDDCM) [13] has later been proposed to eliminate the above mentioned limitations. MDDCM sorts the coefficients in such a way that consecutive coefficients have minimal differences in their magnitude values, and differential values are subsequently multiplied with the input signal.

In order to organize the filter coefficient in an efficient manner, a number of coefficient representation techniques have been proposed in recent times. Pseudo floating point (PFP) is one such scheme which encodes the filter coefficients as sum of signed powers-of-two (SPT) [14]. PFP shifts the first non-zero coefficient and thus considerably reduces the range of the operands in the span part. It results in subsequent minimization of the arithmetic unit and total number of full adders required during the convolution operation. However, PFP-based representation scheme fails to perform better than direct multiplication method when the value of index of first non-zero position equals to zero. In order to further reduce the multiplication cost resulting from PFP-based representation, differential coefficient partitioning algorithm (DCPA) [13] has been introduced. It reduces the range of the span part by partitioning it into two sub-coefficients; where both of them are further scaled by their most significant bit (MSB) and therefore yields drastic reduction in multiplication cost as that of PFP.

In order to eliminate the shortcomings of PFP and DCPA, minimum index floating point (MIFP) representation scheme has been introduced in [15,16], which guarantees reduction in hardware complexity irrespective of the filter coefficients under consideration. This scheme introduces both right and left shift of the span part which eventually minimizes the highest powers-of-two present in a coefficient resulting in a subsequent reduction in the total number of full adder count. It has also been noticed that the efficiency of MIFP scheme may have been deteriorated in case the powers-of-two terms exhibit nonuniform distribution within the coefficient.

This paper introduces the concept of double span floating point (DSFP) representation technique for efficient encoding of powers-of-two coefficients with non-uniform distribution. Proposed algorithm aims at further minimization of the highest powers-of-two term present in any individual coefficient by subdividing the span part into two segments from the highest differential point between two consecutive terms and subsequently MIFP is applied to each of these segments. As a matter of fact, value of each DSFP indices have been reduced significantly with respect to that of MIFP resulting in a satisfactory reduction of full adder count. It has also been observed that for perfectly uniform distribution of coefficient, DSFP is almost as good as MIFP. Proposed DSFP algorithm has been established in this paper with a solid mathematical background and its supremacy over MIFP scheme has also been established. Finally, the performance of DSFP scheme has been compared with that of direct method, PFP and MIFP by including different powers-of-two coefficients with uniform, non-uniform, and random distributions.

2. Proposed double span floating point (DSFP) representation scheme

This section explicitly illustrates the proposed coefficient representation scheme with the help of mathematical backbone and subsequently establishes its superiority over the existing techniques. It has already been reported that MIFP proves to be the most efficient representation scheme for the powers-of-two coefficients of FIR filter in the sense that it requires minimum number of adders during the implementation [16]. Hence, this section has been divided into two segments. Mathematical formulation of the proposed representation scheme has been developed in the first segment, followed by its computational efficiency over MIFP scheme in the next segment.

2.1. Mathematical formulation

Proposed scheme aims at reducing the total numbers of adders even for a non-uniform distribution of the powers-of-two coefficients by dividing the ‘span’ part at n^{th} index such that $|i_k^{n+1} - i_k^n| > |i_k^{j+1} - i_k^j|$ where i_k^j is the j^{th} index of powers-of-two coefficient $h(k)$. As a matter of fact, this highest difference does not appear in the powers-of-two distribution and hence reduces the highest powers-of-two in any coefficient. In case, the span part is not divided into two parts from the highest differential point; total number of full-adder (FA) count would become higher than the optimum one. Each of these two span parts are then organized in such a way that it minimizes the highest index in the respective span part. An individual tap coefficient $h(k)$ under the proposed double span floating point (DSFP) scheme can therefore be represented by:

$$h(k) = 2^{-\mu_{1k}} \sum_{p=1}^{N_{1k}} C_{1p} 2^{-i_{1k}^{p^{DSFP}}} + 2^{-\mu_{2k}} \sum_{p=N_{1k}+1}^{N_k} C_{2p} 2^{-i_{2k}^{p^{DSFP}}} \tag{1}$$

where $\mu_{1k} = i_k^1 + \lceil (i_k^{N_{1k}} - i_k^1)/2 \rceil$ *and* $\mu_{2k} = i_k^{N_{1k}+1} + \lceil (i_k^{N_k} - i_k^{N_{1k}+1})/2 \rceil$

It can be clearly inferred from (1) that the first span involves N_1 number of terms while $N_{2k} = N_k - N_{1k}$ terms are included into the second span. Coefficients C_{1p} *and* C_{2p} inside the summation operation can assume values from the binary set $\mathbb{B} = \{1, -1\}$.

In order to calculate the total number of adders resulting from a given powers-of-two distribution, each of these span parts is taken into account and the required number of adders are computed. Moreover, outputs from these two spans are finally added using a final stage of adders. Total number of one-bit full adders for the proposed DSFP scheme can therefore be written as:

$$\mathbb{T}_{DSFP} = \mathbb{T}_1 + \mathbb{T}_2 + \mathbb{T}_{final}, \tag{2}$$

where $\mathbb{T}_1$ *and* $\mathbb{T}_2$ are the total number of one-bit full adders required for the 1st and 2nd span series. Let N_{1K}^L *and* N_{1K}^R be the length of left and right span of first series and N_{2K}^L *and* N_{2K}^R be the length of left and right span of second series respectively. For an input word length W_x, these adder terms can be represented as:

$$\mathbb{T}_1 = \sum_{p=1}^{N_{1K}^L - 1} \left(i_{1k}^{p_L^{DSFP}+1} + 1 \right) + \sum_{p=1}^{N_{1K}^R - 1} \left(i_{1k}^{p_R^{DSFP}+1} + 1 \right) + W_{1h(k)}^{DSFP} + W_x(N_{1K} - 1) + 1 \tag{3}$$

$$\mathbb{T}_2 = \sum_{p=1}^{N_{2K}^L - 1} \left(i_{2k}^{p_L^{DSFP}+1} + 1 \right) + \sum_{p=1}^{N_{2K}^R - 1} \left(i_{2k}^{p_R^{DSFP}+1} + 1 \right) + W_{2h(k)}^{DSFP} + W_x(N_{2K} - 1) + 1, \tag{4}$$