IJE TRANSACTIONS A: Basics Vol. 33, No. 7, (July 2020) 1223-1230

International Journal of Engineering

Journal Homepage: www.ije.ir

AIOSC: Analytical Integer Word-length Optimization Based on System Characteristics for Recursive Fixed-Point Linear Time Invariant Systems

M. Grailoo, B. Alizadeh*

School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran

PAPER INFO

Paper history:
Received 13 October 2019
Received in revised form 13 October 2019
Accepted 17 June 2020

Keywords:
Analytical Formulation
Fixed-Point Arithmetic
High-Level Synthesis and Optimization
Linear-Time-Invariant System
Range Analysis
Resource Usage

ABSTRACT

The integer word-length optimization known as range analysis (RA) of the fixed-point designs is a challenging problem in high level synthesis and optimization of linear-time-invariant (LTI) systems. The analysis has significant effects on the resource usage, accuracy and efficiency of the final implementation, as well as the optimization time. Conventional methods in recursive LTI systems suffer from inaccurate range estimations due to dependency to symmetry or non-symmetry of the input range over zero, and involvement with parameter adjustments. The under estimations endanger the range safety, and generate a great error due to overflows. On the other hand, the over estimations increase the hardware costs, as well as weaken the signal, if the over estimated ranges are utilized in down-scaling. Therefore, in this paper, we propose an efficient, safe and more precise RA method to measure the range of both recursive and non-recursive fixed-point LTI systems through analytical formulation. Our main idea is to obtain the input sequences for which variables in the LTI system would be maximum and minimum. By applying these sequences to the system, the upper and lower bounds of the intended variables are obtained as the range. The proposed method enhances the bit-widths accuracy more than 34% in average in comparison with the state-of-the-arts. The results also show about 37% and 6% savings in the area and delay, respectively.

***doi**: 10.5829/ije.2020.33.07a.08*

1. INTRODUCTION

The increasing complexity of modern embedded applications in recent decades has forced design methodologies and tools to move to higher abstraction levels. Raising the abstraction levels, and accelerating automation of the synthesis, optimization and verification processes in addition to reducing the time-to-market, help to reduce the verification time as well as facilitate other flows such as accuracy analysis. In high level optimization, a crucial decision to be made is the datapath word length, including the word length of different registers and functional units. In this work, we concentrate on fixed-point representation due to the preference for fixed-point implementations of digital signal processing (DSP) algorithms over floating-point because of hardware cost reduction. Deciding on factors such as integer width and fractional parts of the circuit has significant effects on the resource consumption, accuracy and efficiency of the final implementation. To have finite-precision fixed-point implementations of such systems, range analysis (RA) is an essential and fundamental design step. The analysis characterizes the integer bit-widths (IB) for all the fixed-point variables such that no overflow and underflow occur [1–10].

In this paper, an analytical integer word-length optimization for recursive LTI systems is proposed. LTI systems are the most important category of DSP applications since they include finite-impulse response (FIR), infinite-impulse response (IIR) digital filters, and

* Corresponding Author Email: *b.alizadeh@ut.ac.ir* (B. Alizadeh)

Please cite this article as: M. Grailoo, B. Alizadeh, AIOSC: Analytical Integer Word-length Optimization Based on System Characteristics for Recursive Fixed-Point Linear Time Invariant Systems, International Journal of Engineering (IJE), IJE TRANSACTIONS A: Basics Vol. 33, No. 7, (July 2020) 1223-1230

signal transformations such as Fast Fourier Transform, Discrete Cosine Transform, and Wavelet Transforms [9, 10]. The method not only minimizes the hardware implementation cost, but also reduces the optimization time significantly. In the method, the safe and more precise range is obtained through analytical formulation without any involvement of the parameter adjustments, and without additional iterative operations. The estimations in the method are independent of symmetry or non-symmetry of input range over zero. To do so, the method directly extracts the two input sequences, for which the variables would be maximum and minimum, from the impulse response using the theorem explained in Section 4. The sequences are then applied to the system to obtain the upper and lower bounds of the intended variables. Note that the theorem in this work is also applicable to the feed-forward systems.

The remainder of this paper is organized as follows. Section 2 reviews previous works. Section 3 states our contributions. Section 4 details the proposed range analysis flow through a simple example. Section 5 investigates the experimental results and finally, Section 6 concludes the paper.

2. RELATED WORKS

Several approaches have been introduced to tackle range analysis problems of fixed-point designs which, in general, can be categorized into dynamic and static analyses. Dynamic analysis methods evaluate the system by using input stimulus. This analysis suffers from unsafe, data dependent, and time-consuming estimations, which confine its applicability [8]. Static analysis, however, uses static characteristics of the inputs which are propagated through the system. So, it has recently gained much interest due to safety, no data dependency, and higher efficiency [1–8]. In static analysis, one of the most significant categories is self-validated numerical (SVN) methods. The two most popular SVN methods are interval arithmetic (IA) and affine-arithmetic (AA) [2]. Due to the efficiency of these methods in terms of analysis time, many literatures use them or their extensions to account for RA. The other category of static methods uses more sophisticated approaches such as SMT-based range analysis [7], and hybrid [8] as a combination of IA, AA and AT. These tighter results in the recent addressing methods are obtained at the cost of more analysis time consumption.

Such solutions, however, may not always be adequate, due to being unable to handle recursive circuits, such as IIR filters. Since several fixed-point DSP circuits are based on arithmetic expressions with possible feedbacks, the RA of such circuits, in general, remains still challenging. The main challenge of such systems is to determine final amount of a value when it falls into an infinite loop. In this regard, the methods in [4, 5], utilize L1-norm and L2-norm of impulse responses to compute an inaccurate measurement of the exact range. The L1-norm-based methods in [5] also use the maximum absolute value of the input to obtain the output range. This leads to an over-estimation when the input range is non-symmetric over zero. The over-estimations increase the hardware costs, as well as weaken the signal, if the over-estimated ranges are utilized in down-scaling. The L2-norm-based method in [4] multiplies the maximum absolute value by the L2-norm of the impulse response. The L2-norm-based method under-estimates the ranges when the input is symmetric over zero. The under-estimations endanger the range safety, and generate a great error due to overflows. In order to obtain a tighter range than L1-norm, the method in [3] computes the range by iterative operations of flattening the system, $y[n]$. The analysis will face the problem of adjusting the two parameters to determine the required number of iterations. The parameters are the convergence window size, i.e. w, and the resolution of convergence, i.e. ε. Since the convergence of the algorithm depends on the position of poles and the stability conditions, there is no guarantee to precisely adjust the parameters. So, there is always a probability for an under-estimation in this method which is unacceptable in RA. For comparing our RA method in terms of the precision and hardware cost saving, three methods with the over- and under- estimations are chosen. They include the L1-norm and L2-norm methods due to their prestige and popularity in the scope of analytical range determination of LTI systems. Also, we compare our method with the flattening-based method as an iterative method.

3. OUR CONTRIBUTION

In order to clarify our main contributions, in this section we explain our ideas for efficient RA, obtaining more precise integer bit-widths in a bounded-input, bounded-output (BIBO) stable LTI. Our basic idea in this paper is to analyze the range from the system impulse response without any involvement in any parameter adjustments issues, and iterative operations.

In order to find the output range, we aim to find the input sequences for which the output will be maximum

and minimum. To extract the input sequences, we use the impulse response of a system and the input bounds as will be explained in the following. The maximum and minimum input sequences, as well as the input upper and lower bounds are called $InputSeq_{max}[n]$ and $InputSeq_{min}[n]$, as well as x_{max} and x_{min}, respectively. In the following, we only consider $InputSeq_{max}[n]$, and the primary output variable of y which has the impulse response, i.e. $h[n]$, according to Figure 1(a). Similar arguments exist for intermediate variables with different impulse responses.

The output of a system is obtained by convolving an input sequence and its impulse response. In order to obtain the maximum output, we consider a sequence in a state that has the most overlapping with the impulse response as illustrated in Figure 1(b). In the state, the output maximum is obtained when the input would be in the upper bound, where the impulse response is positive, as well as the input would be in the lower bound, where the impulse response is negative as illustrated in Figure 1(b). This input sequence, which we are looking for to maximize the output, i.e. $InputSeq_{max}$, follows the impulse response form such that places in its input upper, i.e. x_{max}, or lower bounds, i.e. x_{min}, where the impulse response is positive or negative, respectively. Since in the other states with less overlapping, the input sequences generate lower output values, they are not investigated. The $InputSeq_{min}$ is also obtained in a similar way in which the input sequence would be x_{min} and x_{max}, when the impulse response is positive or negative, respectively.

The sequences are then getting backward in time and applied to the system, to account for the output upper and lower bounds. These operations will be repeated for each variable. Since there are variables with the same impulse responses, these variables are grouped together in order to reduce the number of repetitive computations. In fact, the variables, with the same impulse response, constitute a group. Hence, our main contribution is a new method for static RA of LTI systems with or without feedbacks, to achieve safe, more efficient, and more accurate range than the state-of-the-art methods.

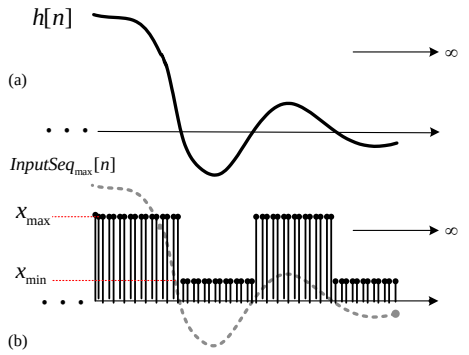

Figure 1. Idea illustration for range analysis of LTI systems: a) impulse response of a system plotted in time domain; b) the input sequence for obtaining upper bound of y when the input is between x_{max} and x_{min}

4. PRPOSED RANGE ANALYSIS

In this section, we propose the RA method, called Analytical Integer Word-length Optimization based on System Characteristics (AIOSC). As mentioned before, RA is crucial for the discrete system design in the implementation of a BIBO stable LTI system. The ranges are used to assign suitable integer bit-widths for all variables such that it is guaranteed that no underflow and overflow happen. Our method finds an input sequence that maximizes the output of a system when it is convolved by the impulse response. The sequence is obtained by following the impulse response form according to Theorem 1. Before introducing the algorithm; we first prove the theorem, which is needed in the rest of this section.

Theorem 1: Two input sequences, in which the BIBO stable LTI system, i.e. $y[n]$, would be maximum and minimum, are $InputSeq_{max}[n]$, and $InputSeq_{min}[n]$, respectively. They are obtained as follows, where $u[n]$ is the unit step function.

$$InputSeq_{max}[n] = x_{max} \times u\big[h[n]\big] + x_{min} \times u\big[-h[n]\big] \quad (1)$$

$$InputSeq_{min}[n] = x_{min} \times u\big[h[n]\big] + x_{max} \times u\big[-h[n]\big] \quad (2)$$

Proof: As discussed in Section 3, the input sequences include only the maximum and minimum of the system input, i.e. x_{max} and x_{min}. Choosing between x_{max} and x_{min} depends on the values of $h[k], k \in \{0,1,\dots,n\}$, as follows:

$$InputSeq_{max}[n] = \begin{cases} x_{max} & if\ h[k] \times x_{max} \geq h[k] \times x_{min} \\ x_{min} & if\ h[k] \times x_{max} < h[k] \times x_{min} \end{cases} \quad (3)$$

$$InputSeq_{min}[n] = \begin{cases} x_{max} & if\ h[k] \times x_{max} \leq h[k] \times x_{min} \\ x_{min} & if\ h[k] \times x_{max} > h[k] \times x_{min} \end{cases} \quad (4)$$

Since $x_{max} \geq x_{min}$, the above relations can be simplified as follows:

$$InputSeq_{max}[n] = \begin{cases} x_{max} & if\ h[k] \geq 0 \\ x_{min} & if\ h[k] < 0 \end{cases} \quad (5)$$

$$InputSeq_{min}[n] = \begin{cases} x_{max} & if\ h[k] \leq 0 \\ x_{min} & if\ h[k] > 0 \end{cases} \quad (6)$$

These relations are equivalent to $InputSeq_{max}[n] = x_{max} \times u[h[n]] + x_{min} \times u[-h[n]]$ and $InputSeq_{min}[n] = x_{min} \times u[h[n]] + x_{max} \times u[-h[n]]$. The sequences can also be obtained through the Equations of (5) and (6).

4. 1. Range Analysis Flow The proposed flow for RA is shown in Figure 2. It takes the input bounds, i.e. $[x_{min}, x_{max}]$, as inputs, and returns the variable integer bit-widths as outputs. This flow is repeated for each group of the variables. In fact, the variables, with the same impulse response, constitute a group in order to reduce the number of repetitive computations. In Step 1, the impulse response for each group is obtained from its linear constant-coefficient difference equation (LCCDE), if it currently does not exist. In Step 2, the input sequences, i.e. $InputSeq_{max}[n]$ and $InputSeq_{min}[n]$, are found based on Theorem 1. In this step, the function $UnitStep()$ from Mathematica is invoked to apply the unit step function to the impulse response. In Step 3, these sequences are getting backward in time, and applied to the system. This response can be obtained by direct evaluation of the convolution sum of the sequences and the impulse response, as indicated in the figure where "$*$" denotes convolution. However, since the convolution in the time domain corresponds to multiplication in the z-domain, another simple alternative is obtaining the response in the z-domain. So, the z-transform of the sequences, and the impulse response can be created by the function $ZTransform()$ from Mathematica [11]. Then the z-transform of the impulse response is multiplied by the z-transforms of the sequences. Finally, the function $InverseZTransform()$ is invoked to obtain the corresponding results in the time domain, i.e. the $UpperBound\,[n]$ and $LowerBound\,[n]$. In Step 4, the minimum and maximum of the mentioned functions (called a and b) will constitute the final range, i.e. $[a, b]$. To obtain the bit-width including sign bit from the range, the following relation is employed.

$$i = \lceil log_2(max\,(|a|, |b|)) \rceil + \alpha,$$
$$\alpha = \begin{cases} 1 & if\ \ mod(log_2(b)) \neq 0 \\ 2 & if\ \ mod(log_2(b)) = 0 \end{cases} \quad (7)$$

4. 2. Example In order to clarify the flow, let us consider the example of $y[n] = \alpha y[n-1] + \beta x[n]$, with $\alpha = 0.8$ and $\beta = 0.5$. The example is a low pass filter, which enjoys wide applications in control systems, Kalman filtering, communication processing to reduce noise, and image averaging. The filter with all the input and intermediate variables, as some vertical rectangles, is shown in Figure 3(a). In this example, the variables x_2 to x_4 offer the same impulse response, which differs from the impulse response of x_1. So the variables are broken down into two groups: x_1 in G_1, and x_2 to x_4 in G_2. For G1, first (according to Step 1) the impulse response $h[n]$ is obtained by using direct and inverse z-transform. In order

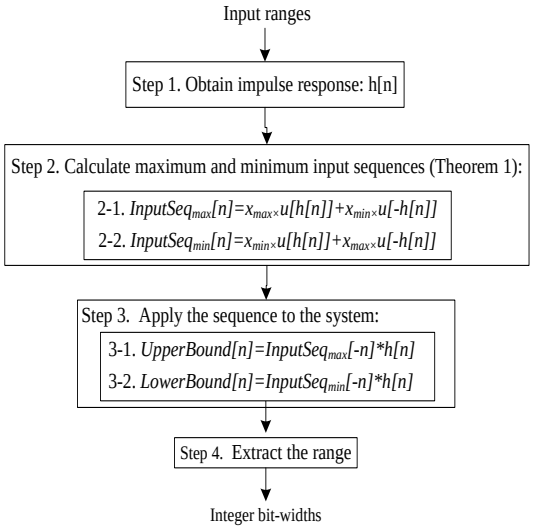

Figure 2. Proposed range analysis flow

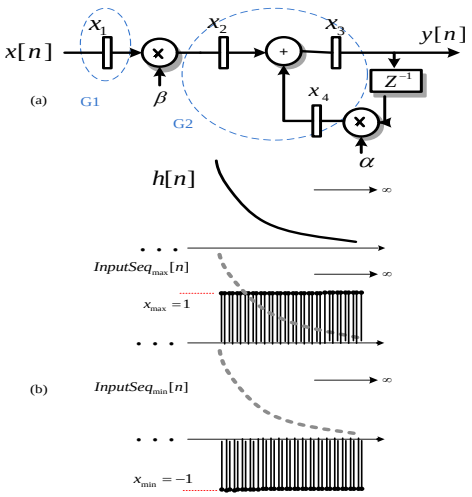

Figure 3. a) one-pole digital filter with intermediate variables; b) the impulse response and input sequences, $InputSeq_{max}$ and $InputSeq_{min}$, of the filter

to obtain $h[n]$, the other way is to solve the difference equation of $y[n]$, when $x[n]$ is replaced by $\delta[n]$, and $y[n]$ is replaced by $h[n]$. The impulse response for the variables in G_1 would be $h[n] = 0.5\ (0.8)^n u[n]$. Second, based on Theorem 1, the input sequences for $x_{max} = 1$ and $x_{min} = -1$ would be $InputSeq_{max}[n] = u[n]$ and $InputSeq_{min}[n] = -u[n]$. The impulse response of $h[n]$ and the input sequences of $InputSeq_{max}[n]$ and $InputSeq_{min}[n]$ are depicted in Figure 3(b). As shown in this figure, $InputSeq_{max}[n]$ gets the maximum input when $h[n]$ is positive, and the minimum input when $h[n]$ is negative. Since in this case, $h[n]$ is always positive, $InputSeq_{max}[n]$ would be $u[n]$, and vice versa for $InputSeq_{min}[n]$. Third, since the backward of the sequences in time domain are also unit step functions, these sequences are applied to the system as follows:

$$UpperBound[n] = \mathcal{Z}^{-1}\left\{\frac{0.5z}{z-0.8} \times \mathcal{Z}\{u[n]\}\right\} = \mathcal{Z}^{-1}\left\{\frac{0.5z}{z-0.8} \times \frac{z}{z-1}\right\} = 2.5 - 2 \times e^{-0.223144n} \quad (8)$$

$$LowerBound[n] = \mathcal{Z}^{-1}\left\{\frac{0.5z}{z-0.8} \times \mathcal{Z}\{-u[n]\}\right\} = \mathcal{Z}^{-1}\left\{\frac{0.5z}{z-0.8} \times \frac{-z}{z-1}\right\} = -2.5 + 2 \times e^{-0.223144n} \quad (9)$$

Finally, when n approaches infinity, the range x_3 is obtained which is $[-2.5,2.5]$. These values for x_1, x_2, and x_4 are $[-1,1]$, $[-0.5,0.5]$ and $[-2,2]$, respectively. The integer bit-widths for x_1 to x_4 are $IB_{x_1} = 2$, $IB_{x_2} = 1$, $IB_{x_3} = 3$ and $IB_{x_4} = 3$. The obtained output range and integer bit-width by L2-norm for x_3 are $[-0.83,0.83]$ and $IB_{x_3} = 1$, respectively. It is obvious that these measurements under-estimate the exact ones.

5. EXPERIMENTAL RESULTS

In order to demonstrate the applicability of our proposed method in different types and forms of the recursive LTI systems, as well as the superiority of the method over the state-of-the-arts, we have provided several benchmarks with various forms and types. The forms are direct (DR), parallel (PRL), and cascade (CS), as well as the types are high-pass (HPF), low-pass (LPF), and band-pass (BPF) filters. Bench #3 is a bi-quad eighth-order cascaded structure of four 2nd-order direct-form IIR filters. The last benchmark is also a National Television Systems Committee (NTSC) channel cascaded eighth-order LPF IIR filter with the cutoff frequency of 4.74MHz. The details of the benchmarks such as type, order, numerator, and denominator coefficients are given in Table 1. Our algorithm has been implemented with Mathematica, and run on an Intel 4702MQ core i7 with 8 GBs of main memory, running Linux operating system. For the synthesis process, the tool Xilinx ISE V14.1 on the Virtex-7 FPGAs target has been chosen. The device contains user-programmable elements known as slices, dedicated multiply-and-add units, DSP blocks and embedded RAMs. In order to make fair comparisons, the designs are implemented by using slices and combinatorial elements without any pipelining. The variable indexes in the feedback parts have been numbered in a clockwise direction. In the first experiment, we compare AIOSC with L2-norm-based method (L2-norm) in [4] and flattening-based methods in [3] to show the precision of our method. It is assumed that the primary inputs are symmetric over zero, and lie within the normalized range of $[-1,1]$. The estimated range, bit-widths, and their under-estimation ratio have been reported in Table 2. In the table, the first major column has listed the benchmarks. The second and third major columns include the estimated ranges and bit-widths by the RA methods. Finally, the last column shows the under-estimation ratio of the AIOSC than the state-of-the-art methods. As shown in the table, L2-norm when the

TABLE 1. Range and bit-width evaluation results of AIOSC and L2-norm for the primary output variable

Bench #	Estimated Range			Estimated bit-width			Underestimation Ratio %			
	AIOSC	L2-norm	Flattening-based	AIOSC	L2-norm	Flattening-based	L2-norm/AIOSC		Flattening-based /AIOSC	
							Range	Bit	Range	Bit
1	[-270.89,270.89]	[-103.33,103.33]	[-232.15,232.15]	10	8	9	162	25	17	12
2	[-4.58,4.58]	[-2.18,2.18]	[-4.47,4.47]	4	3	4	109	33	3	0
3 Quad	[-76.23,76.23]	[-25.40,25.40]	[-75.25,75.25]	8	6	8	200	33	2	0
4 NTSC	[-275.12,275.12]	[-73.26,73.26]	[-273.92,273.92]	10	8	10	275	25	5	0
				Average underestimation ratio %			186.5	29	6.75	3

TABLE 2. Benchmark features

Bench#	Type	Form	Order	Numerator Full-precision Coefficient	Denominator Full-precision Coefficient
1	**HPF**	**DR**	2	101.8, -203.4, 101.6	1, -1.967, 0.968
2	**LPF**	**PRL**	3	2,0.1,-0.4	1,0.1,-0.46,0.08
3 Quad	**BPF**	**CS**	8	1, 2, 1 1, -2, 1 1, 2, 1 1, -2, 1	1, a, b 1, -a, b 1, c, d 1, -c, d a=0.47583613785934908, b=0.63399428536347535, c=1.0921588046377746, d=0.87447915380668007 1, a, b 1, c, d 1, e, f 1, g, h
4 NTSC	**LPF**	**CS**	8	1,2,1 1,2,1 1,2,1 1,2,1	a=-0.7093449002973562, b=0.19225253081578914, c=-0.22413592126247239, d=0.41113157239125847, e=0.27362911645488941, f=0.66517393946636161, g=0.57030039990570558, h=0.88861236005184185

input is symmetric over zero under-estimates ranges and bit-widths, in all benchmarks. The under-estimations are more in the higher order benchmarks of Quad and NTSC. The ranges and bit-widths under-estimations are about 186% and 29% on average, respectively. Hence the estimations generate a great error due to overflows. Obtaining the exact output range requires the exhaustive simulations by feeding all possible sequences into inputs. The sequences are infinite for recursive filters. So, generating all possible infinite sequences is time consuming and even impossible in high order filters.

In the flattening-based method, the window size and the resolution are considered $(w, \varepsilon) = (10,1)$. As illustrated in the table, the ranges are under-estimated in all benchmarks. The under-estimations in the first benchmark lead to the under-estimated bit-width. In the other benchmarks, if the under-estimated ranges are used in the signal down-scaling, it can cause the overflow in the variables, which encompass the larger numbers. Let us consider the second benchmark. The flattening-based method estimates the maximum absolute range of 4.47 while the output variable can accept the number ± 4.58. In this case, all signals are divided by 4.47 and the output encompasses the number 1.02. The number is more than one which led to the output overflow. So, the flattening-based method under-estimates range and generates a great error.

In the next experiment, we concentrate on the safe methods, i.e. L1-norm-based method (L1-norm) [5], in comparison to AIOSC. In the experiment, it is assumed that the primary inputs are non-symmetric over zero and lie within the range of [9,10]. The bit-widths estimations for primary outputs are depicted in Figure 4. In this figure, the other estimations include the ranges plus the improvements are also shown as some entries of the small tables beside the bit-width bars. As seen in this figure, the L1-norm-based method constitutes over-estimations when the input bound is not symmetric over zero. The range over-estimations in some benchmarks are more than 20 times than the estimated range by AIOSC. If the over-estimated ranges are utilized in down-scaling, the range can strongly weaken the signal. Moreover, the range over-estimations result in an additional integer bit for the all benchmarks. As seen, by increasing the range over-estimations, the excess bits are also growing. The excess bits growing have significant impact on hardware cost. The amount of the impact is investigated in the next experiment. In the experiment, the AIOSC method shows the bit-width improvement of about 34.75% on average.

As mentioned, the effect of inaccuracy in the opposite direction, i.e., over-estimations instead of under-estimations, is on the hardware cost. Whatever the ranges of all intermediate and output variables are more exact, we expect to achieve the smallest bit-widths, leading to a

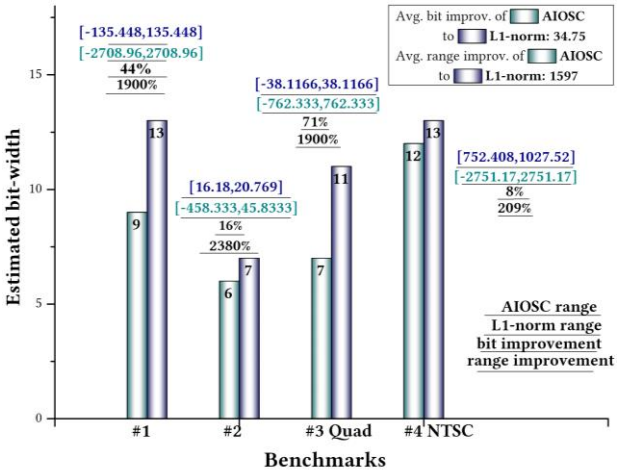

Figure 4. The estimation results of different RA methods

reduction in the circuit area and delay. In order to complete the experiment, the effect of the over-estimations on the area and delay are investigated in Figure 5. This figure indicates the area costs of the benchmarks for the assigned bit-widths obtained by AIOSC and L1-norm when using Xilinx ISE for the synthesis process. In this figure, other results of delay plus area and delay savings are also shown as some entries of the small tables beside the area bars. As illustrated in Figure 5, area and delay almost follow the estimated bit-widths. It means, in the positions that one method has estimated lower bit-widths; the delay and area are pursuing this flow and become less. The area (slice) and delay saving of AIOSC is 37.25% and 5.6% in comparison with L1-norm, respectively.

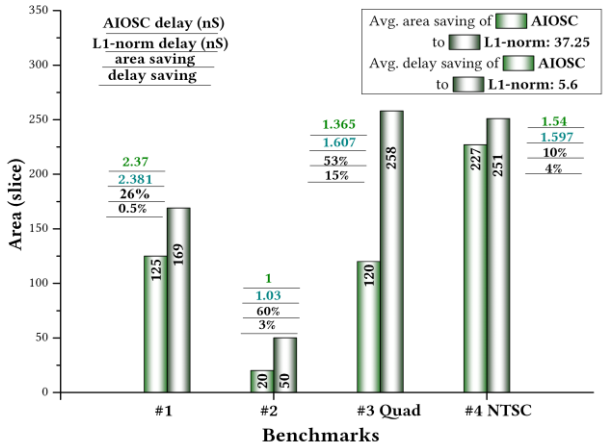

Figure 5. Area and delay comparison of RA methods

6. CONCLUSIONS AND FUTURE WORK

The range analysis plays an important role in high-level synthesis of arithmetic circuits, as it can directly impact the overall design cost and performance. Most of existing analyses on the recursive LTI systems estimate the bounds inaccurately. It leads to produce some great errors or increase the hardware cost. Therefore, in this paper, a new, more accurate and efficient RA method for fixed-point recursive LTI systems was proposed. The method obtained a safe and more precise range and bit-width estimations from the impulse response, without any involvement of the parameter adjustments, and without any additional iterative operations. The proposed method brought advantages of 29% bit-width improvement. It led to 37.25% and 5.6% area and delay saving in comparison with the previous state-of-the-art methods.

As our future work, we are going to extend our method to support the error analysis in LTI systems with feedback for the maximum mismatch (MM), mean square error (MSE) and signal to quantization noise ratio (SQRT) metrics.

7. REFRENCES

1. Grailoo, M., Alizadeh, B. and Forouzandeh, B., "Improved range analysis in fixed-point polynomial data-path", ***IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems***, Vol. 36, No. 11, (2017), 1925–1929. doi: 10.1109/TCAD.2017.2666607
2. Grailoo, M., Alizadeh, B. and Forouzandeh, B., "UAFEA: Unified analytical framework for IA/AA-based error analysis of fixed-point polynomial specifications", ***IEEE Transactions on Circuits and Systems II: Express Briefs***, Vol. 63, No. 10, (2016), 994–998. doi: 10.1109/TCSII.2016.2539078
3. Sarbishei, O., Radecka, K., and Zilic, Z., "Analytical optimization of bit-widths in fixed-point LTI systems", ***IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems***, Vol. 31, No. 3, (2012), 343–355. doi: 10.1109/TCAD.2011.2170988
4. Abbas, M., Gustafsson, O. and Johansson, H., "On the fixed-point implementation of fractional-delay filters based on the Farrow structure", ***IEEE Transactions on Circuits and Systems I,*** Vol. 60, No. 4, (2013), 926–937. doi: 10.1109/TCSI.2013.2244272
5. Rocher, R., Menard, D., Scalart, P. and Sentieys, O., "Analytical Approach for Numerical Accuracy Estimation of Fixed-Point Systems Based on Smooth Operations", ***IEEE Transactions on Circuits and Systems I***, Vol. 59, No. 10, (2012), 2326–2339. doi: 10.1109/TCSI.2012.2188938ï
6. Chung, J. and Kim, L. W., "Bit-Width Optimization by Divide-and-Conquer for Fixed-Point Digital Signal Processing Systems", ***IEEE Transactions on Computers***, Vol. 64, No. 11, (2015), 3091–3101. doi: 10.1109/TC.2015.2394469
7. Kinsman, A. B. and Nicolici, N., "Bit-width allocation for hardware accelerators for scientific computing using SAT-modulo theory", ***IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems***, Vol. 29, No. 3, (2010), 405–413. doi: 10.1109/TCAD.2010.2041839
8. Pang, Y., Radecka, K., and Zilic, Z., "An efficient hybrid engine to perform range analysis and allocate integer bit-widths for arithmetic circuits", In Proceedings of the Asia and South Pacific Design Automation Conference, ASP-DAC, (2011), 455–460. doi: 10.1109/ASPDAC.2011.5722233
9. Chernoyarov, O.V., Golpaiegani, L.A., Glushkov, A.N., Lintvinenko, V.P. and Matveev, B.V., "Digital binary phase-shift keyed signal detector", ***International Journal of Engineering - Transactions A: Basics***, Vol. 32, No. 4, (2019), 510–518. doi: 10.5829/IJE.2019.32.04A.08
10. Kumar, P., and Kumar Chaudhary, S., "Stability and Robust Performance Analysis of Fractional Order Controller over Conventional Controller Design", ***International Journal of Engineering - Transactions B: Applications*** , Vol. 31, No. 2, (2018), 322–330. doi: 10.5829/ije.2018.31.02b.17
11. Wolframe Mathematica: definitive system for modern technical. [Online]. Available: http://www.wolfram.com/mathematica/

Persian Abstract

چکیده

بهینه‌سازی طول کلمه صحیح یا تحلیل دامنه، یک مساله چالش‌برانگیز در بهینه‌سازی و سنتز سطح بالای سیستم‌های خطی نامتغیر با زمان بازگشتی محسوب می‌گردد. این تحلیل، تاثیر بسزایی بر مصرف منابع، دقت، کارایی و زمان بهینه‌سازی می‌گذارد. روش‌های پیشین، از معایبی چون تخمین نادقیق دامنه شامل تخمین مازاد یا تخمین زیر مقدار واقعی رنج می‌برند. این عدم دقت بدلیل وابستگی روش‌ها به تقارن ورودی نسبت به صفر و همچنین وابستگی به برخی از پارامترها می‌باشد. تخمین‌های زیر مقدار واقعی، سبب ایجاد سرریز و تولید خطاهای بزرگ می‌شود. از طرف دیگر، تخمین مازاد، هزینه سخت افزار را افزایش می‌دهد. همچنین اگر این تخمین مازاد در مقیاس کردن استفاده شود، سبب تضعیف سیگنال می‌گردد. بنابراین در این مقاله یک روش تحلیل دامنه دقیق، کارا و ایمن با روش‌های تحلیلی برای اندازه‌گیری دامنه در سیستم‌های خطی نامتغیر با زمان بازگشتی و غیربازگشتی برای طراحی‌های ممیز ثابت، ارائه می‌شود. ایده اصلی یافتن دنباله ورودی برای هر متغیر است که به ازای آن خروجی سیستم ماکسیسمم و مینیمم گردد. با اعمال این دنباله‌ها به سیستم، محدوده بالا و پایین هر متغیر به عنوان دامنه بدست می‌آید. روش ارائه شده، دقت طول کلمه را بطور متوسط تا بیش از ۳٤٪ در مقایسه با روش‌های قبلی بهبود می‌بخشد. نتایج همچنین ۳۷٪ بهبود در مساحت و ٦٪ بهبود تاخیر را نشان می‌دهد.