2016 IEEE Computer Society Annual Symposium on VLSI

High-Accuracy FIR Filter Design using Stochastic Computing

Bo Yuan
Department of Electrical Engineering
City University of New York, City college
New York City, NY 10031
byuan@ccny.cuny.edu

Yanzhi Wang
Department of Electrical Engineering Computer Science
Syracuse University
Syracuse, NY 13244
ywang393@syr.edu

***Abstract*—Finite impulse response (FIR) filter is the basic functional component in various signal processing and communication systems. In many practical applications that have stringent requirement on spectrum, long FIR filters are needed to achieve the desired filtering performance. However, because a *T*-tap FIR filter requires *T* copies of high-complexity multiplier, the conventional design of long FIR filter consumes a large amount of silicon area and power dissipation. This paper, for the first time, proposes a high-accuracy stochastic computing (SC)-based FIR filter design. By utilizing the simplicity of stochastic arithmetic unit, the proposed stochastic FIR filter achieves significant reduction in hardware complexity as compared to the conventional design. More importantly, this paper proposes a new high-accuracy non-scaled stochastic adder that has significant increase in computation accuracy than the conventional stochastic adder. Built on this new stochastic adder, the proposed stochastic FIR filter achieves much higher accuracy than the existing stochastic FIR filter design, especially for large *T* cases, thereby unlocking the potentiality for the widespread applications of stochastic FIR filters in practical signal processing systems.**

Keywords—stochastic computing; FIR filter; high accuracy

I. INTRODUCTION

Filter is the fundamental spectrum-shaping unit in the modern digital signal processing, image processing and communication systems. In general, based on the different requirements in the design specification, a targeted spectrum-shaping behavior can be implemented by either a finite impulse response (FIR) filter or an infinite impulse response (IIR) filter [1]. In particular, FIR filter is usually preferred because of its liner phase response and non-recursive property, which makes FIR filter has stronger resilience to phase shift and higher stability in long-term processing than its IIR counterpart. As a result, to date FIR filters are widely adopted in numerous commercial signal processing products, especially in the baseband processor market.

However, despite its unique advantage on robustness, the FIR filter falls short in hardware performance as compared to IIR filter. This is mainly because FIR filter requires higher order than the IIR filter to achieve the same spectral performance, thereby causing the increasing need of taps of FIR filter. Particularly, in many practical applications such as optical communications, long FIR filters have to be adopted to meet the stringent requirement in design specification. As a result, the corresponding hardware performance is very inferior since this type of long FIR filters consumes a large amount of high-complexity multipliers.

To address the aforementioned high complexity problem of long FIR filter design, [2] proposed to utilized *stochastic computing* (SC) [3-4] as the underlying number presentation method. Different from the conventional 2's complement *binary computing* (BC), SC utilizes a stream of bits to represent a number, where the number is the ratio of bit "1" over the entire stream. Such probabilistic-theory-based representation has unique advantage on low-cost hardware and high-speed clock rate. However, a straightforward stochastic FIR filter design suffers from severe accuracy loss when the tap of FIR (denoted as *T*) increases due to the inherent down-scaling property of stochastic adder. Although [2] developed a new stochastic inner product approach to alleviate this down-scaling effect, the approach proposed in [2] has very limited scalability. As a result, to date the efficient hardware design of long stochastic FIR filter for practical applications is still very challenging.

This paper, *for the first time*, proposed a high-accuracy stochastic FIR filter. By leveraging a new type of SC representation, a high-accuracy stochastic adder, which achieves significant increase in accuracy over conventional stochastic adder, is developed. More importantly, the proposed stochastic adder is non-scaled, thereby completely avoiding the original down-scaling effect for large *T* cases. Based on the proposed stochastic adder, the high-accuracy stochastic FIR is developed. Performance analysis shows that the proposed stochastic FIR achieves very high accuracy and very low hardware cost, thereby paving the way of the widespread application of stochastic FIR filter in various practical applications.

The rest of this paper is organized as below. Section II gives a brief review of the stochastic computing and FIR filter theory. Section III first analyzes the challenge of the long stochastic FIR design, and then develops the proposed high-accuracy non-scaled stochastic adder. The hardware architecture of the stochastic FIR filter is also presented in this section. Section IV discusses and analyzes the performance of the proposed stochastic FIR in terms of computation accuracy and hardware performance. The conclusions are drawn in Section V.

978-1-4673-9039-2/16 $31.00 © 2016 IEEE
DOI 10.1109/ISVLSI.2016.63

IEEE computer society

II. Background

A. Stochastic Computing (SC)

Stochastic computing is an emerging computing technique that is well-suited for low-complexity design. Deviated from conventional binary computing, the stochastic computing interprets the value with a stream of bits. Here the value of the represented number is a function of the ratio of bit "1" over the entire bit-stream. For instance, a number *x* within range [0, 1] can be represented by a length-*N* bit-stream containing *X* bits "1", where *x*=*X*/*N*. Note that because here each bit in the bit-stream has the same weight, the same number can be interpreted by different bit-streams (see Fig. 1(a)). In [3-4], this straightforward use of the ratio of "1" was denoted as *unipolar* representation since only positive bounded value can be interpreted. Fig. 1(b) shows the corresponding hardware design for the multiplication using this unipolar representation. It is seen that only one AND gate is needed to implement the stochastic multiplication, thereby significantly reducing the hardware complexity as compared to the original binary-computing-based multiplier.

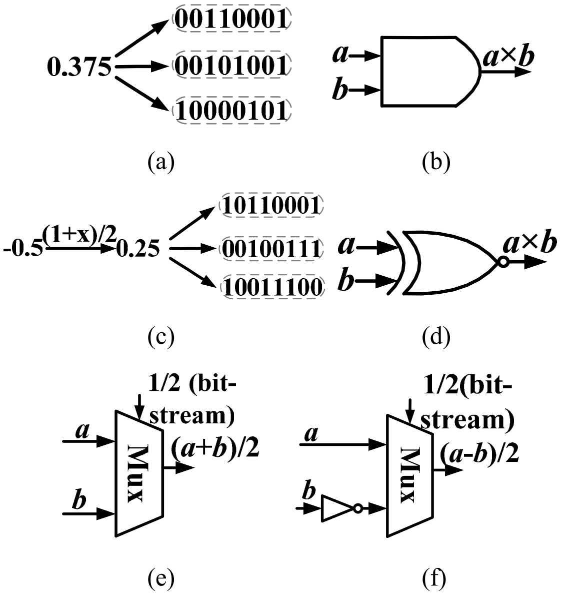


Fig. 1. (a) Example unipolar representation (b) Unipolar multiplication (c) Example bipolar representation (d) Bipolar multiplication (e) Scaled addition (f) Scaled subtraction. Cited from [5].

One drawback of the unipolar representation is its positive-only limitation. Therefore, in [3-4] a *bipolar* representation was proposed to include the negative number into the framework of stochastic computing. More specifically, for a number *x* within range [-1, 1], its bipolar representation is *x*=2(*X*/*N*)-1, where *X* and *N* is the number of bit "1" and length of stream, respectively. Fig. 1(c) illustrates the bipolar representation of an example -0.5 by different bit-streams. It should be noted that since the underlying representation has changed, the stochastic multiplication unit using bipolar representation (see Fig. 1(d)) is different from that using unipolar representation.

Besides its simplicity on multiplication, stochastic computing also enables significant reduction in the implementation for other basic arithmetic units. Fig. 1(e) and (f) show the stochastic circuits for the down-scaled addition and the down-scaled subtraction, respectively. Notice that here both the unipolar and bipolar representations use these two implementations. For the other types of stochastic arithmetic unit, the reader is referred to [3-4] [6-7].

B. Finite Impulse Response (FIR) Filter and stochasitc FIR

As indicated in Section I, FIR filter is a fundamental component in digital and image signal processing systems. In general, the output sequence of a *T*-tap FIR filter can be calculated as below:

$$y[n] = h_0x[n]+h_1x[n-1]+\ldots+h_{T-1}x[n-T+1]$$
$$= \sum_{i=0}^{T-1} h_i x[n-i], \qquad (1)$$

where *x*[*n*] is the input signal, *y*[*n*] is the output signal, *T* is the tap of the filter and h_i is the coefficient of the filter.

From (1) it is seen that a *T*-tap FIR filter is involved with *T* times of multiplication for each output sample. Therefore, in many practical applications where large *T* is needed in the design specification, the overall area cost and power consumption of long FIT filter are very significant due to the high hardware complexity of multiplier. To overcome this problem, stochastic computing is a promising solution for low-cost FIR design since the multiplication unit in SC is very simple. In [8-9], both the direct-format and lattice-format stochastic FIR filters were developed. However, as analyzed in Section III-A, the existing stochastic FIR filter designs suffer from accuracy loss problem, especially for large *T* cases, thereby hindering the application of stochastic FIR filters in practical systems.

III. High-Accuracy Stochastic FIR Fitlter

A. Challenge of Long Stochastic FIR Filter Design

To date, several stochastic FIR filter designs had been reported in [2] [8-9]. Next, we show that these designs are not suitable for long FIR filters due to the severe accuracy loss problem.

1) Direct-Format Stochastic FIR Filter Direct-format design is the straighforward implementation of (1). Fig. 2 shows the overall architecture of the direct-format stocahstic FIR fitler. Here SA is the stocahstic adder with inner architecture as Fig. 1(e), and SM is the stochastic multiplier with inner architecture either as Fig. 1(b) (unipolar represetnation) or Fig. 1(d) (bipolar represetnation).

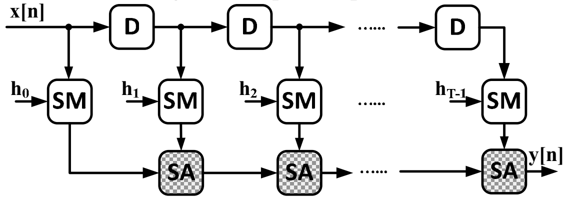


Fig. 2. Direct-form stochastic FIR.

Notice that because all the stochastic adders (SA) are the down-scaled version in Fig. 1(e), the actual output in Fig. 2 is actually the down-scaled of the desired output *y*[*n*]. In other words, if we denote the actual output in Fig. 2 as $y_{scale}[n]$, then $y_{scale}[n]=y[n]/2^{T-1}$. Such down-scaling phenomenon causes severe accuracy loss when T is large. This is because the precision of the stochastic computing system using length-2^L bit-stream can only achieve as low as $1/2^L$, which is insufficient to represent the $y_{scale}[n]$ that requires precision as $1/2^{T-1}$ when T is large. Even worse, in many cases when $y[n]$ is already very small, the $y_{scale}[n]$ in Fig. 2 is just zero due to the insufficient representation precision. Notice that such imprecision cannot be compensated by the post-processing (up-scaling) of $y_{scale}[n]$ since the information has already be lost during the down-scaling procedure. Tap-wise up-scaling for each intermediate sum in each SA is a potential solution to compensate down-scaling immediately, however, the finite state machine (FSM)-based stochastic linear gain unit [4] that implements the up-scaling operation consumes a large amount of hardware as well as introducing extra inaccuracy. As a result, such in-time compensation approach is inefficient for long stochastic FIR filter either.

2) Lattice-Format Stochastic FIR Filter Lattice-format design is another popular implementaiton of FIR filter. Fig. 3 shows the overall architecture of the lattice-format stochastic FIR filter. Here the k_i is the transformatted coefficient for the lattice format, and it can be calcualted from h_i [1].

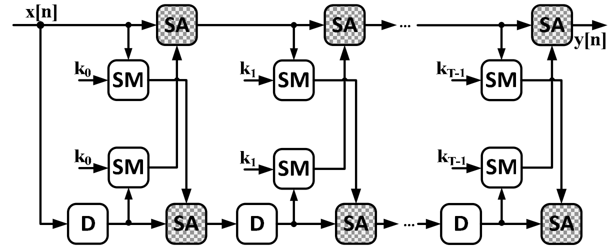


Fig. 3. Lattice-format stochastic FIR.

Similar to Fig. 2, the actual result in Fig. 3 is also the down-scaled version $y_{scale}[n]=y[n]/2^{T-1}$ due to the use of down-scaled adder in Fig. 1(e). Even worse, as indicated in [8-9], the architecture in Fig. 3 requires extra binary-to-stochastic (B-to-S) and stochastic-to-binary (S-to-B) conversion units around each D flip-flops. Because the B-to-S and S-to-B units consume a large amount of area and power, the use of lattice-format stochastic FIR is inefficient in neither accuracy nor hardware performance.

3) Inner-Product-based Stochastic FIR To address the down-scaled problem of stochastic FIR filter, [2] proposed a reformulated stoahstic inner product (see Fig. 4). Different from conventional stochastic inner product, the selection signal of the multiplexor in Fig. 4 is the uneven weighted sum instead of 1/2. As a result, the scaling factor in Fig. 4 is $1/(|b_0|+|b_1|)$ instead of 1/2. Notice that since (1) can be viewed as the inner product of two size-T vectors, the stochastic FIR filter can be alternatively implemented via a size-T stochastic inner product unit consisting of multiple size-2 stochastic inner product units in Fig. 4 where the scaling factor is $1/(|h_0|+|h_1|+\ldots+|h_{T-1}|)$. Compared to the direct-format design in Fig. 2 with scaling factor as $1/2^{T-1}$, the inner-product-based stocahstic FIR fitler has improved computing accuracty since down-scaled effect is allivated. Notice that the lattic-format stochastic FIR filter in Fig. 3 can also be reformulated via using the uneven-weight inner product in Fig. 4 to improve its computation accuracy.

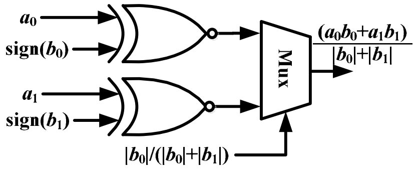


Fig. 4. Uneven-weight –based stochastic size-2 inner product. Cited from [5].

Although the uneven-weight inner product can alleviate the down-scaled problem in some cases, the severe accuracy loss still occurs for long FIR filter. This is because $|h_0|+|h_1|+\ldots+|h_{T-1}|$ can still be very large when T is large, or even T is a medium value with large h_i. As shown in Section IV, the computation accuracy of the inner-product-based stochastic FIR filter decrease rapidly as T increases. Moreover, the use of Fig. 4 is based on the assumption that the coefficient h_i is pre-known. However, in some practical applications those coefficients need to be adjusted according to design specifications, thereby greatly limiting the generality of the inner-product-based approach.

To sum up, despite the prior efforts on stochastic FIR filter, the efficient design of high-accuracy long stochastic FIR filter for practical applications is still a challenging problem.

B. High-Accuracy Non-Scaled Stochastic FIR Filter

In this subsection, we propose a high-accuracy stochastic FIR filter for arbitrary T. First, we develop a high-accuracy non-scaled stochastic adder as the key component. Then, the overall architecture of the FIR filter is presented.

1) High-Accuracy Stochastic Adder Different from unipolar-based or bipolar-bsaed down-scaled stochastic adder, the proposed high-accuracy stochatic adder is based on a two-line SC represetnation [10]. Here this two-line representation is introduced first.

a) Two-line SC representation Fig. 5 illustrates the two-line SC representation for x= -0.5. Here $M(X_i)$ and $S(X_i)$ are the i-th bits of the *magnitude stream* and *sign stream*, respectively. In general, if the length of the stream is 2^L, then these two bit-streams jointly represent x as below:

$$x=\frac{1}{2^L}\sum_{i=0}^{2^L-1}(1-2S(X_i)M(X_i)) \quad (2)$$

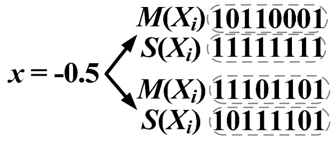


Fig. 5. Example two-line SC representation.

From (2) it can be seen that, different from unipolar or bipolar representation, the two-line SC representation introduces the concept of "+1" and "-1" for calculating the represented number. As discussed in next subsubsection, this difference further leads to different architecture of the stochastic arithmetic unit.

Similar to the SC system using unipolar or bipolar representation, the two-line-based SC also needs a binary-to-stochastic (B-to-S) conversion unit. Fig. 6 shows the overall architecture for the B-to-S unit that contains a random number generator (RNG) and a comparator. Here for each 2's complement number x, each bit in its sign stream is its own sign bit. Notice that this phenomenon only exists when x is converted to two-line bit-stream, while in the SC system each bit in the sign stream is not necessarily the same.

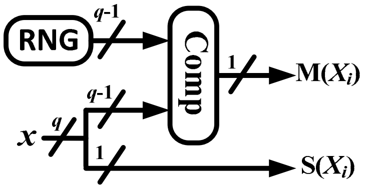


Fig. 6. B-to-S unit for two-line SC representation.

b) High-Accuracy Non-Scaled Stochastic Adder Design: Based on the two-line SC reprentation, we propose a high-accuracy non-scaled stochastic adder. The key idea of the proposed adder is based on the following observation: In (2) the i-th bits of magnitued stream and sign stream jointly contribute $(1\text{-}2S(X_i))M(X_i)$ to x. If we deonte A_i=$(1\text{-}2S(A_i))M(A_i)$ and B_i=$(1\text{-}2S(B_i))M(B_i)$ as the i-th contriubtion for the two inputs a and b, respectively, then the i-th contribution of their sum c=a+b, referred as C_i, should be the element of {1,0,-1} plus the carry bit. As a result, if we utilize a three-state counter to store the positive or negative carry bit, then we can get the following truthtable for C_i:

TABLE I. TRUTH TABLE FOR C_I

$A_i + B_i$	Current Counter	Next Counter	C_i
2	-1	0	1
2	0	1	1
2	1	1	1
1	Remains	Remains	1
0	-1	0	-1
0	0	Remains	0
0	1	0	1
-1	Remains	Remains	-1
-2	-1	-1	-1
-2	0	-1	-1
-2	1	0	-1

Notice that here since C_i is the contribution of the i-th bits of magnitude and sign bit-streams, we have $(1\text{-}2S(C_i))M(C_i)$=C_i. Accordingly, $S(C_i)$ and $M(C_i)$ can be determined as below:

TABLE II. GENERATATION SCHEME FOR OUTPUT STREAMS

C_i	$S(C_i)$	$M(C_i)$
1	0	1
0	0 or 1	0
-1	1	1

Based on Table I and Table II, the hardware architecture of the proposed non-scaled adder is shown as below:

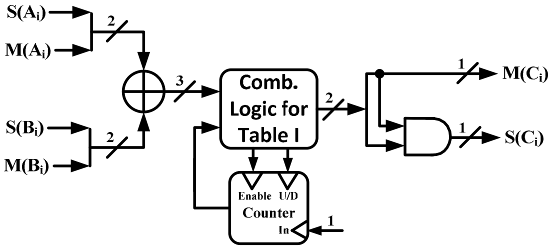
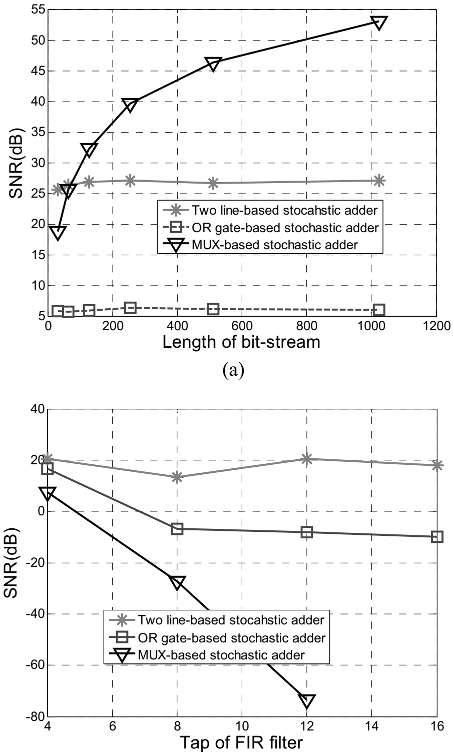


Fig. 7. Hardware architecture of the proposed stochastic adder.

Compared to the conventional stochastic adder in Fig. 1(e), the proposed adder in Fig. 7 has two advantages. First, it is non-scaled adder. Although OR gate can be used as approximated non-scaled adder in some applications [3], the OR-gate-based adder suffers from huge approximation error as well as limitation for positive-only addition, thereby causing the OR-gate-based adder has very severe accuracy loss if we consider the negative inputs (see Fig. 8(a)). Second, the proposed stochastic adder has very high accuracy in term of signal-to-noise ratio (SNR), especially for the adder chain case. As shown in Fig. 8, although the mux-based adder has better computation accuracy than the proposed two-line-based design for single adder case (see Fig. 8(a)), its performance degrades significantly if the application (such as FIR filter) needs the use of adder chain (see Fig. 8(b)), while the computation accuracy of the proposed two-line-based design still remains at a relatively high level.



(a)

(b)

Fig. 8. SNR of different stochastic adders for (a) single adder case. (b) adder chain. Here the length of bit-stream is 256.

2) High-Accuracy Stochastic FIR Filter: Based on the proposed stochastic adder in Fig. 9, the high-accuracy non-scaled stochastic FIR filter can now be developed since stochastic adder was the compoenent that introduces down-scaled effect. Notice that because here two-line SC represtiation is adopted, the implemetantion of stocahstic multiplication (see Fig. 9) is different from that in Fig. 1(b) or (d).

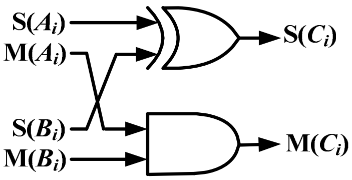


Fig. 9. Two-line stochastic multiplier. Cited from [11].

With the use of the two-line-based non-scaled stochastic adder and multiplier, the hardware architecture of the high-accuracy stochastic FIR filter can be developed. According to the design specification, the overall architecture can be chosen as either direct format or lattice format. Fig. 10 shows the overall architecture of the proposed stochastic FIR filter in a direct format. Here TSM and TSA denote two-line SM and two-line SA, respectively. Notice that the propogated values ($x[n-i]$) among D flip-flops are in the 2's complement form to reduce the uncessary use of delay components.

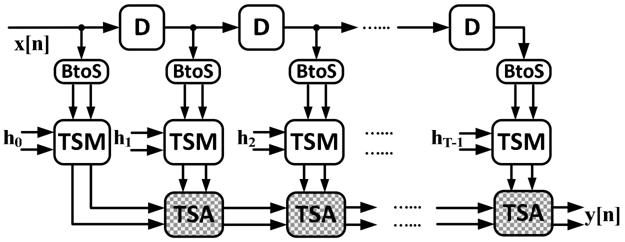


Fig. 10. The proposed direct-form at non-scaled stochastic FIR filter.

Besides B-to-S conversion units, a complete stochastic computing system also requires S-to-B conversion units as the interface between stochastic computing and binary computing domains. Fig. 11 shows the hardware architecture of the S-to-B conversion units in the scenario of two-line representation. Here different from the S-to-B units in bipolar or unipolar case, an additional AND gate is required to convert the two bit-streams to one bit-stream, followed by a regular up/down counter to calculate the final results.

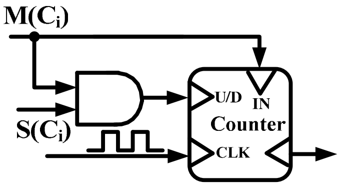


Fig. 11. S-to-B unit for two-line SC representation.

IV. Performance and Analysis

A. Computation Accuracy

As analyzed in Section III-A, the accuracy loss for large T is the most challenging problem for stochastic FIR filter designs. Fig. 12 compares the computation accuracy of different types of stochastic FIR filter designs. Here the even weighted-based design indicates the FIR filter adopts multiplexor as the stochastic adder. As it is seen from this figure, compared to the state-of-the-art stochastic FIR filter designs, the proposed two-line-based design shows significant advantage on computation accuracy. In particular, the proposed stochastic filter is the only design that can achieve beyond 10dB SNR performance in the long tap region, thereby enabling it is well-suited for practical applications that need long FIR filters.

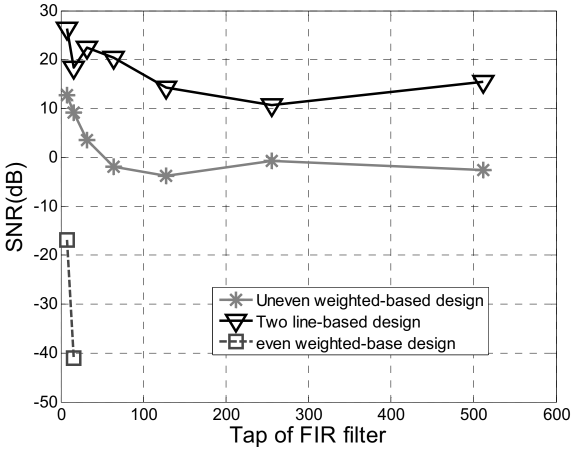


Fig. 12. Computation accuracy for different stochastic FIR filters. Here the length of bit-stream is 256.

B. Hardware Perforamnce

Table III lists the estimated hardware performance of the different FIR filter designs. Here the tap of the filter T is set as 64. In addition, the data-width for the 2's complement-based design is 8, which corresponds to length-256 bit-stream for the uneven-weight design in [2] and the proposed two-line design in this paper.

TABLE III. Estimated Hardware Performance of FIR Filters

Design	Two-line	Uneven-weight [2]	2's complement
CMOS Tech	90nm		
Area(um^2)	44800	18304	76800
Clock Frequency(MHz)	750	800	400

From Table III it can be seen that, the proposed two-line-based design achieves 41% reduction in silicon area as compared to the conventional 2's complement design. Although the proposed design has higher hardware complexity than the uneven-weight design in [2], the two-line-based design has much better computation accuracy in large T cases, which is the main working region for the practical long FIR filters. Consequently, the proposed two-line stochastic FIR filter is the first FIR filter that can achieve both low hardware complexity and sufficient computation accuracy.

V. Conclusion

This paper presents high-accuracy area-efficient stochastic FIR filter design. With the use of two-line stochastic computing representation, a high-accuracy non-scaled stochastic adder is developed. Based on this stochastic adder, the non-scaled high-accuracy stochastic FIR filter is presented. Analysis show that the proposed stochastic FIR filter achieves both low hardware cost and high computation accuracy, thereby it is very suitable for practical systems.

References

[1] A. Y. Oppenheim, R. W. Schafer, J. R. Buck, et al., Discrete-time signal processing, vol. 2. Prentice-hall Englewood Cliffs, 1989.

[2] Y-N. Chang and K. K. Parhi, "Architectures for digital filters using stochastic computing," in Proc. of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),pp. 2697-2701, 2013.

[3] B. Gaines, “Stochastic computing systems,” Advances in Information Systems Science, vol. 2, no. 2, pp. 37–172, 1969.

[4] B. Brown and H. Card, “Stochastic neural computation I: computational elements,” IEEE Trans. Comput., vol. 50, no. 9, pp. 891-905, Sept. 2001.

[5] B. Yuan, Y. Wang and Z. Wang, "Area-Efficient Error-Resilient Discrete Fourier Transformation Design using Stochastic Computing," accepted by ACM 26th Great Lakes VLSI Symposium (GLSVLSI’2016)

[6] P. Li, D. J. Lilja, W. Qian, and K. Bazargan, “Computation on stochastic bit streams: Digital image processing case studies,” IEEE Trans. on Very Large Scale Integrated (VLSI) Systems, vol. 22, no. 3, pp. 449-462, April 2013.

[7] W. Qian, X. Li, Marc D. Riedel, K. Bazargan, and D. J. Lilja, “An architecture for fault-tolerant computation with stochastic logic,” IEEE Trans. on Computers, vol. 60, no. 1, pp. 93-105, 2011.

[8] K. K. Parhi and Y. Liu, “Architectures for IIR digital filters using stochastic computing,” in Proc. of IEEE International Symposium on Circuits and Systems (ISCAS), pp. 373-376, June 2014

[9] Y. Liu and K. K. Parhi, “Lattice FIR digital filters using stochastic computing,” in Proc. of 2015 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), pp. 1027-1031, April 2015

[10] S. L. Toral, J. M. Quero, and L. G. Franquelo, “Stochastic pulse coded arithmetic,” in Proc. of IEEE Int. Symp. Circuits Syst. (ISCAS), pp. 599–602, May 2000.

[11] J. Yang, C. Zhang, S. Xu and X. You, “Efficient stocahstic detector for large-scale MIMO,” accepted by IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)