Design and implementation of DA FIR filter for bio-inspired computing architecture

B. U. V. Prashanth¹, Mohammed Riyaz Ahmed², Manjunath R. Kounte³
¹,³School of Electronics and Communication Engineering, REVA University, India
²School of Multidisciplinary Studies, REVA University, India

ABSTRACT
This paper elucidates the system construct of DA-FIR filter optimized for design of distributed arithmetic (DA) finite impulse response (FIR) filter and is based on architecture with tightly coupled co-processor based data processing units. With a series of look-up-table (LUT) accesses in order to emulate multiply and accumulate operations the constructed DA based FIR filter is implemented on FPGA. The very high speed integrated circuit hardware description language (VHDL) is used implement the proposed filter and the design is verified using simulation. This paper discusses two optimization algorithms and resulting optimizations are incorporated into LUT layer and architecture extractions. The proposed method offers an optimized design in the form of filters average minimizations of the number of LUT, reduction in populated slices and gate minimization for DA-finite impulse response filter. This research paves a direction towards development of bio inspired computing architectures developed without logically intensive operations, obtaining the desired specifications with respect to performance, timing, and reliability.

This is an open access article under the CC BY-SA license.

1. INTRODUCTION
High throughput is required since the finite impulse response filters are used intensively in video, communications systems as well as bio-inspired computing systems. Essentially, digital filters are used in time and frequency domain to adjust the characteristics of the signals and are identified as the primary digital signal processing feature [1]. The DSP design techniques focus mainly on multiplier-based architectures for multiply-and-accumulate (MAC) blocks implementation which represent the FIR filters and several functions. High speed parallel filter designs are elucidated in excruciating detail. Finite impulse response (FIR) filters are prominent building blocks for several applications in the field of digital signal processing (DSP). High-speed FIR filters have been widely used to perform signal equalization on the received data in real time due to the increasing demand for video-signal processing and transmission. Therefore a structured VLSI architecture is needed for a programmable fast FIR filter [2].

The various FIR Filters were suggested in last few decades, many structures and different algorithms have been utilized for the enhancement of the filter weights. The very common structures utilized were least mean square (LMS) derived models since their response in convergence is strong. Block processing with distributed arithmetic methods is explored to derive a design that should give high throughput [3]. The
Distributed arithmetic (DA) is a strategy of high-speed multiplication which is a bit serial word parallel technique where the throughput rate does not depend on the data size. The DA facilitates to avoid the multipliers in the design and makes the area of the system efficient in the throughput and several DA based structures were designed in order to minimize the area and to reduce the cost of processing [4]. The primary operations necessary for DA-based processing are a series of accesses to a lookup table (LUT), preceded by the LUT output's shift-accumulation operations. The standard framework of DA used to implement the FIR filter implies that the coefficients of the impulse response are fixed and this action allows use of ROM based LUTs. However, with linear filter order the memory requirement for Distributed Arithmetic implementation of FIR filters rises exponentially is one of the hard problems to be addressed [5]. The key contributions of this research are:

- Develop systolic array architecture with tightly coupled co-processor based data processing units.
- Develop optimization algorithms with optimizations incorporated into LUT layer with architecture extractions and propose bio inspired computing architecture to compute FIR filters at high processing speeds using reconfigurable computing based on DA strategy.

2. RELATED WORK

Modular finite-impulse response (FIR) filter whose filter coefficients switch dynamically during latency, which plays a major role in architectures for software-defined radio (SDR), multi-channel filters, bio-inspired computing and digital up/down converters. However, when the filter coefficients vary dynamically, the well-known multiple constant multiplication (MCM)-based methods that are widely used to realize the FIR filters cannot be used. Addressing to the solution to the problem of such large memory requirement, systolic decomposition techniques are utilized for DA-based implementation of long-length convolutions and FIR filter of large orders. It is necessary to use rewritable RAM based LUT instead of ROM based LUT for reconfigurable DA based FIR filter whose filter coefficients alter dynamically. Another method is to store the analog domain coefficients using serial digital to analog converters, resulting in mixed-signal architectures [6].

A pipelined design for an adaptive FIR filter carry out the save accumulation technique which is used for partial inner product calculation that facilitates in enhancing the throughput with block processing is utilized in increasing the computational speed of the system. On the other hand, a particular multiplier-based structure requires a wide chip region, and thereby controls limitations on the highest allowable order of the filter that can be interpreted for high-throughput applications [7]. In recent years, distributed arithmetic (DA)-based technique has gained substantial popularity due to its high capacity for processing throughput and increased regularity, resulting in cost-effective and area-time efficient computing structures.

The primary operations required for DA-based processing are a sequence of accesses to a lookup table (LUT), followed by the LUT output's shift-accumulation operations [8]. The conventional implementation of the DA used to implement the FIR filter assumes that the coefficients of the impulse response are fixed and this behavior allows the use of ROM based LUTs. However, with the filter order the memory requirement for DA-based implementation of FIR filters increases exponentially [9].

The systolic decomposition techniques are used to get rid of the problem of such a large memory requirement. For long-length convolutions and large-order FIR filter for DA-based implementation, we must use rewritable RAM based LUT instead of ROM based LUT for reconfigurable DA-based FIR filter whose filter coefficients change dynamically. Another approach is to store the coefficients in the analog domain by using serial digital to analog converters resulting in mixed-signal architecture. We also find quite a few works on DA based implementation of adaptive filters, where the coefficients change at every cycle [10].

3. PROPOSED METHOD AND ALGORITHM DESIGN

Distributed arithmetic is a popular architecture without the use of multipliers to implement FIR filters. DA makes efficient use of LUTs, shifters, and adders to calculate the sum of products required for FIR filters. Since these operations effectively map onto an FPGA, Distributed arithmetic on these devices is a favourable architecture [11].

The Figure 1 illustrates the experimental design of the research work presented in this manuscript. Distributed Arithmetic is a prominent architecture without the use of multipliers to implement FIR filters. DA makes efficient use of LUTs, shifters, and adders to calculate the sum of multiplication factors needed for FIR filters. Though distributed arithmetic implements the FIR filter by serialization bits of inputs, a filter quantisation is required. Due to the fixed data path requirements in input analog to digital converter (ADC)
and the output digital to analog converter (DAC) widths the length of the word with 12 bit input and output with 11 fractional bits are assumed to be required to quantize the FIR filter [12].

After the quantization process the HDL Code is generated with DA architecture. The HDL code generator uses distributed arithmetic architecture, and partitions the look-up-table (LUT) into a specified number of LUT partitions with the range of taps each partition associates. It is best to divide the taps into a number of LUTs for a filter with many taps, with each LUT storing the sum of coefficients only for the taps that are associated with it.

The FIR filter structure has symmetric coefficients, and we consider converting the structure to reduce the area. Here we convert the filter structure to direct form symmetric and generate the HDL code for default radix of 2. In hardware, a symmetrical filter structure offers advantages, as it halves the number of coefficients to work with which substantially reduces the complexity of the hardware. The predefined architecture is an implementation of Radix 2 that runs on one bit of input data per clock period. Before an output is obtained, the number of clock phases elapsed is equal to the number of bits in the input data and DA may effectively limit the throughput. DA can be configured to process multiple bits in parallel, to improve the DA throughput. The processing of 12 bits at a time for a 12 bit input word length can be specified with the corresponding DA-Radix values of $2^{12}$. The speed vs. area is trade off by selecting different 'DARadix' values and the amount of parallel bits illustrates the factor with the increased rate of the clock which is the number of cycles to perform an iteration [13]. The Tables 1-3 elucidate the information of DA architecture. The Table 1 depicts the 'DARadix' values with corresponding values of number of cycles to perform an iteration and multiple for LUT sets for the given filter. Further Table 2 illustrates the details of LUTs with corresponding 'DALUTPartition' values. Details of LUT indicate number of LUTs with the sizes of LUT for example (1x1024x18) implies 1 LUT of 1024 18-bit wide locations [14].

Table 1. 'DARadix' values with number of cycles to perform an iteration and multiple for LUT sets

<table>
<thead>
<tr>
<th>Folding Factor</th>
<th>LUT Sets Multiple</th>
<th>DA Radix</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>$2^{12}$</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>$2^6$</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>$2^4$</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>$2^3$</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>$2^2$</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>$2^1$</td>
</tr>
</tbody>
</table>
Table 2. Details of LUTs with corresponding partitions of DA-LUT

<table>
<thead>
<tr>
<th>Address Width</th>
<th>Size (bits)</th>
<th>LUT Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>b’259072</td>
<td>[1204x13], [13x4096], [14x4096], [15x4096], [18x4096]</td>
<td>(12) (12) (12) (10)</td>
</tr>
<tr>
<td>b’147544</td>
<td>[2x2048x14], [1x2048x18], [2x2048x13], [1x8x11]</td>
<td>(11) (11) (11) (11) (3)</td>
</tr>
<tr>
<td>b’78080</td>
<td>[1x1024x16], [3x1024x15], [1x1024x16], [1x256x13]</td>
<td>(10) (10) (10) (10) (8)</td>
</tr>
<tr>
<td>b’43712</td>
<td>[1x512x12], [1x512x14], [1x512x18], [1x6x12], [2x512x13], [1x512x15]</td>
<td>(9) (9) (9) (9) (9) (4)</td>
</tr>
<tr>
<td>b’25384</td>
<td>[1x4x10], [1x256x14], [4x256x13], [1x4x16]</td>
<td>(8) (8) (8) (8) (8) (2)</td>
</tr>
<tr>
<td>b’14248</td>
<td>[1x128x14], [1x128x16], [1x4x10], [1x128x12], [3x128x13], [1x128x18]</td>
<td>(7) (7) (7) (7) (7) (2)</td>
</tr>
<tr>
<td>b’1412</td>
<td>[4x6x12], [1x6x12], [1x6x17], [2x6x14], [1x6x16], [1x6x13]</td>
<td>one(1,1)*(6), (4)</td>
</tr>
<tr>
<td>b’4696</td>
<td>[1x8x11], [4x32x12], [3x32x13], [1x32x15], [1x32x14], [1x32x17], [1x32x11]</td>
<td>one(1,1)*(5), (3)</td>
</tr>
<tr>
<td>b’2904</td>
<td>[1x16x15], [5x16x12], [2x16x13], [1x16x17], [1x4x10], [3x16x11]</td>
<td>one(1,4)*(4), (2)</td>
</tr>
<tr>
<td>b’1926</td>
<td>[8x8x12], [2x8x14], [2x8x15], [1x8x17], [1x2x7], [5x8x11], [1x8x13]</td>
<td>one(1,9)*(3), 1</td>
</tr>
<tr>
<td>b’1412</td>
<td>[12x4x11], [2x4x13], [2x4x15], [1x4x17], [2x4x10], [6x4x12], [4x4x14]</td>
<td>one(1,29)*(2)</td>
</tr>
</tbody>
</table>

As depicted in Table 3, if it is required to increase the clock rate by four scales the sampling frequency and utilize six input LUTs then we can verify that the details of LUT meets the area requirements. Next a test bench is designed with a standard setup, and uses a simulator to verify the generated code for distributed arithmetic architecture [15]. The synthesis tool is utilized to compare the area and speed of the DA architecture. The Algorithm 1 presents the performance analysis and optimization of LUT layer. As shown in Algorithm 2, the cost function could be any arbitrary parameters delay, power or power delay multiplication (PDM) returned from optimized LUT.

Algorithm 1: Performance analysis and optimization of LUT layer

Result: Optimization of LUT Layer

Start
Optimize LUT (Addr bits: k, num LUTs: m)

\[
\text{Delay}(LUT_i, 1) \leftarrow d_{lut}[j] \quad \text{for all } j \text{ set of } [k];
\]

\[
\text{Power}(LUT_i, 1) \leftarrow p_{lut}[j] \quad \text{for all } j \text{ set of } [k];
\]

\[
\text{Power Delay}(LUT_j, 1) \leftarrow p_d_{lut}[j] \quad \text{for all } j \text{ set of } [k];
\]

While {Read the Input Parameters}{

\[
\text{for } (i=2; i \leq k; i++)
\]

\[
\text{for } (j=2; j \leq m; j++)
\]

\[
\text{else If}{\text{Perform Optimization}}
\]

\[
\text{Delay}(LUT_i, j) \leftarrow \text{minu}\{\text{max}\{ d_{lut}[u] + D(LUT_i-u, j-1)\) \};
\]

\[
\text{Power}(LUT_i, j) \leftarrow \text{minima w}\{p_{lut}[w] + \text{Power}(LUT_i-w, j-1)\};
\]

\[
\text{PDM}(LUT_i, j) \leftarrow \text{minima} \{\text{Delay}(LUT_i, j) \cdot \text{Power Utilization } (LUT_i, j)\};
\]

\[
\text{end for}
\]

\[
\text{Compute: return Delay(LUT_k,m), Power(LUT_k,m), Power Delay(LUT_k,m)}
\]

Calculate the performance;

} Stop

Algorithm 2: Algorithm steps to optimized architecture extractions

Result: Optimized Architecture Extractions

Start;

Define parameters;

While {Read the Input Parameters}{

Architecture Optimize (N:Filter Order):
Design and implementation of DA FIR filter for bio-inspired computing architecture (B. U. V. Prashanth)

4. RESULTS AND DISCUSSIONS

The fixed point settings are applied in order to obtain the characteristic plot of magnitude response (dB) indicating the curves between the magnitude (dB) and the normalized frequency (π radians per sample) with the comparison between reference and quantized filter as depicted in Figure 2(a). The characteristic plot representing the complete design specification of DA FIR filter along with the Log magnitude (dB) and phase (degrees) is as depicted in Figure 2(b). In this case the full precision override is not considered and custom coefficient data type is considered in the design. With the optimizations addressed by variations in architectural level enhancements using DA concept of digital filtering which improves device utilization [16, 17].

![Figure 2. Plot of (a) magnitude response (dB), (b) log magnitude (dB)-phase (degrees)](image)

Here the clock rate is four times the input sample rate for this architecture and the effective filter length for serial partition value is 58 along with three samples of HDL latency, achieved with the FIR compiler and the corresponding frequency response diagram obtained in FIR compiler is as depicted in Figure 3(a) and with reference to this the pole-zero (P-Z) diagram is as depicted in Figure 3(b). Because of mid-stage pipelining, the entire architecture is split into two sections, namely the input section and the output section. Here the power consumption of the DA architecture is estimated at 20 MHz frequency and the final DA architecture is designed using the systolic rearrangement of delay elements. The preconfigured logic functions, that is the intellectual property (IP) cores optimized for FPGAs is generated using FIR compiler and Figure 4 illustrates the block design to verify the DA FIR filter responses as obtained in the Figure 2 and Figure 3.
Figure 3. Plot of (a) frequency response (dB), (b) pole-zero (P-Z) diagram

As depicted in Figure 4 the RAM based shift register is having 16 bit width and 16 bit depth is configured as a circular buffer and it is initialized with memory initialization radix and memory initialization vector of 16-bits as arbitrary waveform generator and on every cycle of 100 MHz clock, the shift RAM outputs the last sample first and proceeds towards the initial sequence and loops back. Further the complete DA FIR filter is processed using the ZYNQ FPGA as a special purpose tightly coupled processor. The Figure 5 illustrates the performance evaluation of the design with behavioral simulation of DA FIR Filter obtained in Xilinx ISE environment with phase (phase 0, 3) and serial (serial out 1, 2, 3) and the Figure 6 depicts the performance evaluation with analysis of filter coefficient values.

The Figure 7 compares the proposed DA FIR filter design with the previous designs available in [18-21] in terms of number of multipliers versus the filter order as depicted in Figure 7(a). In Figure 7(b) the number of adders versus order of Filter is illustrated along with the LUT optimization with number of LUTs versus order of filter in Figure 7(c), the Figure 7(d) represents the number of registers versus the filter order. The estimated delay based on (Gate delay-DG) for Distributed Arithmetic unit of LUT, LUT-less and
proposed architecture implementation [22, 23] is as shown in the Figure 8. Here the delay of the proposed architecture is 14% (for 8-order filter) and 64.7% (for 140-order filter) less delay in comparison of LUT-less architecture [24, 25].

Figure 5. Performance evaluation with simulation of DA-FIR Filter with phase, and serial outputs

Figure 6. Performance evaluation with simulation of DA-FIR Filter with filter coefficients

Figure 7. Comparisons of previous research with proposed research with filter order versus (a) Number of multipliers, (b) Number of adders, (c) Number of LUT’s, (d) Number of registers
The Table 4 illustrates the comparison of the obtained synthesis results of the proposed research work with the previously published literature. Here the factor B is denoted as bit width of the filter coefficients. Further extending this concept in the following mathematical formulae compares the hardware and time complexities of our proposed DA-FIR structures with other filter structures.

Here if: 
- \( B \): = bit width of filter coefficients,  
- \( M \): = image width,  
- \( T_M \): = delay of multiplier,  
- \( TPA \): = delay of parallel adder.

Then: 
- \( T_1 = T_{PA} + 2 \cdot T_{FA} \cdot (\log_2 N - 1) \),  
- \( T_{DAHLUT} = 2T_{PA} \),  
- \( T_{PU} = T_{MUX} + T_{SAB} \),  
- \( T_{SAB} = T_{FA} + T_{XOR} + T_D \)

The \( T_{MUX}, T_{FA}, T_{XOR} \) and \( T_D \) are the delay of MUX, full adder, XOR gate and D flip-flop, respectively. This comparison of time complexities and hardware of proposed DA-FIR designs with other filter designs is as depicted in Table 5. The Table 4 illustrates our best solution and compares the obtained parameters of our synthesis results with previous works in terms of numerical values of (MSP)-Minimum Sampling Period(ns), Area(\( \mu m^2 \)), Power(mW), (ADP)-Area Delay Product(\( \mu m^2 \) ns), Energy per output, Throughput(MHz) [29, 30]. Further the Table 5 compares the obtained results in our work with previous works with numerically addressing with mathematical formulas of various parameters such as Throughput, multipliers, adders and registers [31, 32]. The implementation of multi-core computing system is done on the ZYNQ platform with the use of VERILOG language to program and compile the framework [33].
5. CONCLUSIONS

The VHDL is used implement the proposed DA finite impulse response filter and the design is verified using simulation. The calculated theoretical values of the design match with obtained practical values in the real time simulation environment. Two optimization algorithms are proposed and the resulting optimizations are incorporated into LUT layer and architecture extractions of designed block. The proposed work offers an optimized design in the form of average reductions of number of LUT, reduction in populated slices and reduction in the number of gates for DA-finite impulse response filter implementation. This research paves a way for bio inspired computing architecture with reconfigurable computing strategies designed to avoid computationally intensive operations, achieving the desired specifications with respect to flexibility, timing, and performance.

REFERENCES


Table 5. Time complexities and hardware of proposed DA-FIR designs with other designs

<table>
<thead>
<tr>
<th>Design</th>
<th>Throughput</th>
<th>Multipliers</th>
<th>Adders</th>
<th>Registers</th>
</tr>
</thead>
<tbody>
<tr>
<td>[34]</td>
<td>{L(TM + T1)}</td>
<td>M^2</td>
<td>M^2−1</td>
<td>(M+N)(M−1)</td>
</tr>
<tr>
<td>[35]</td>
<td>{l(TM+2T2)}</td>
<td>MN^2</td>
<td>L[(N+1)−1]</td>
<td>(M+N)(M−1)</td>
</tr>
<tr>
<td>This Work</td>
<td>N((3TPx + 2.B.Tm)</td>
<td>-</td>
<td>N((L+N)−2b) + L(3M^2/5)−1</td>
<td>N(M−1)+N.Lc</td>
</tr>
</tbody>
</table>

Design and implementation of DA FIR filter for bio-inspired computing architecture (B. U. V. Prashanth)


