A Survey on Fast Convolution Circuit with Weighted Partitioning

Pratibha Singh[1], Prof. Rishi Jha[2]

[1]*Mtech Student,* [2]*Guide*

Department of Electronics and Communication Engg. NIIST, Bhopal

Abstract- Multiplying a variable by an arrangement of known constant coefficients is a typical task in numerous igital signal processing (DSP) approaches. Compared to other common operations in DSP algorithms, such as addition, subtraction, using delay elements, etc., multiplication is generally the most expensive. In order to achieve higher computational throughput and/or lower energy consumption, many DSP algorithms are implemented in custom hardware. A field– programmable gate arrays (FPGAs) and application specific integrated circuits (ASICs) are few examples of custom hardware. These benefits are further enhanced by finding lower cost implementations of constant coefficient multiplication, as smaller and cheaper implementations of DSP systems are obtained. In ASICs, a smaller logic circuit results in less material costs and lower energy consumption. Alternatively, for the same amount of silicon, a higher computational throughput can be obtained by placing more computational units in parallel. This examination presents an extensive survey of literature on Multiplierless Multiple-Constant Convolution Circuit.

Keywords- Constant Convolution Circuit, Multiplierless Circuit, FPGA, Multiplierless Multiplication, Digital Filter, VLSI

I. INTRODUCTION

In recent years, with the advance of semiconductor manufacturing technology, the requirements of digital very-large-scale-integrated (VLSI) circuits, which are composed of tens to hundreds of millions of gates, have led to many challenges during manufacturing test. Moreover, the unprecedented levels of design complexity and the gigahertz range of operating frequencies make the testing of nanometre system-on-chip (SOC) designs a most demanding challenge. This is because the large and complex chips require a huge amount of test data and dissipate a substantial amount of power during test, which greatly increases the system cost. There are many test parameters that should be improved in order to reduce the test cost. These parameters include the test power, test length (test application time), test fault coverage, and test hardware area overhead.

In the mid-1980‟s, field-programmable gate arrays were introduced in the commercial market and thereon, have been produced en masse. The design of digital hardware has undergone rapid changes in the meanwhile. They have become indispensable system parts because of their versatility to execute different logic functions and easy reconfigurablity in step with changing hardware requirements.

The easy availability of logic, routing resources has allowed for the existence of hardware with hundreds of FPGA devices. This has been useful for verifying prototype logic designs and parallel running computing platforms. But a matter of concern remains that the leaps and bounds in hardware prowess are not reflected in the marketed software for mapping designs automatically.

Most users agree that the greatest restriction frequently experienced in the use of contemporary FPGA systems is the average time taken to place and route circuits inside a single board. Typically in FPGAs, the above operations can take hours and in some cases, days compared to tens of minutes on other options. On board implementation isn‟t necessarily a simple task as conceptual reality is impeded by the above restriction. If the developer is using resources to develop one digital design over several weeks/months, then compilation period in the order of hours is acceptable but if only computing is required of the resources, it is otherwise. In

FPGAs, a bargain has to be reached when compared with ASIC platforms. When set against semi custom platforms, FPGAs can offer lower prices because of mass production. Design mapping is also a moot point. For a fixed size device, placement and routing based on the picked algorithm depends on the efficiency of optimization. Optimality in VLSI is measured in terms of chip size, wire length, delay minimization etc. which have a direct impact on the manufacturing cost, the IC performance and its power consumption.

Scaling of transistor geometries have led to integration of millions of devices in a very small space, thus driving realization of complex applications on hardware and supporting high speed applications. While the basic principles are largely the same, the design practices have changed enormously because of the increases in and transistor budgets and clock speeds, the growing challenges of power consumption and the improvements in productivity and design tools. Device scaling has increased the operating frequency of many applications, but has led to high power consumption. This incredible growth has

come from steady miniaturization of transistors and improvements in manufacturing processes. Most other fields of engineering involve tradeoffs between performance, power and price. However, as transistor become smaller, they also become faster, dissipate less power and are cheaper to manufacture. This synergy has revolutionized not only electronics, but also industry at large. In order to reduce power, many researchers, designers and engineers have come up with many innovative techniques and have patented their ideas. Nevertheless, designers will need to budget and plan for power dissipation as a factor nearly as important as performance and perhaps more important than area.

An alternative architecture for performing convolution is a dedicated VLSI processor where the convolution operation is built-in the silicon structure. Consequently, there is no complex instruction decoding and large cache module. The architecture of the processor is optimised only for the convolution operation. The typical structure of the VLSI processor is given in Figure 1.1, and incorporates multipliers and adders to perform arithmetic operations, and delay elements to feed the arithmetic block with proper data.

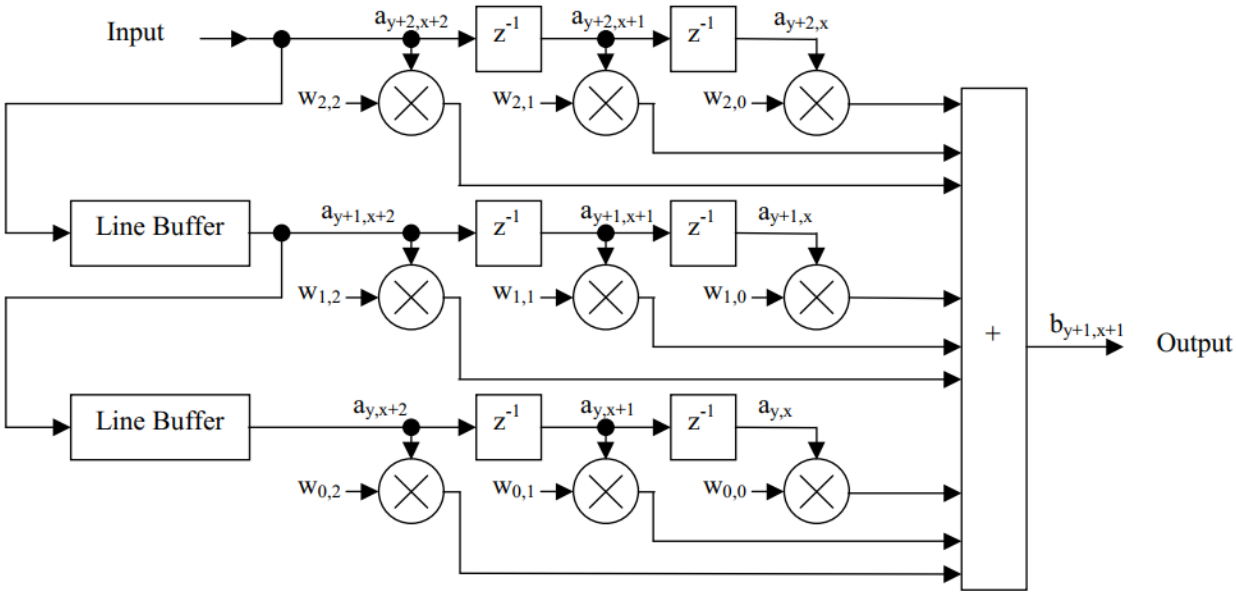

Fig. 1.1 VLSI architecture for 3X3 convolution.

II. MULTIPLIERLESS MULTIPLICATION (MM)

Multiplication is a fundamental operation performed in the convolution. The way a multiplication is carried out in ASIC and FPGA designs initially seems to be very similar. Both ASICs and FPGAs require the same algorithms to be implemented. For example the structure of parallel-array multipliers or Wallace tree multipliers for FPGAs and ASICs are very similar. Nevertheless, the most important advantage of FPGAs over ASICs is reconfiguration which allows for a change of the multiplication coefficient either by the change of the multiplicand (an input to a fully functional multiplier denoted further as Variable Coefficient Multiplier VCM) or by the change of a Constant Coefficient Multiplier circuit. The Constant Coefficient Multiplier, in comparison to the VCM, has much lower hardware requirements, and therefore is recommended providing that a coefficient value is relatively constant during the calculation process.

FPGAs implement logic cells as a Look Up Table (LUT) memory, therefore the inherent way of performing multiplication seems the LUT based Multiplication (LM) for which large LUT memory is split and combined with adders. Conversely, ASIC solutions usually implement the Constant Coefficient Multiplier employing a Multiplierless Multiplication (MM), where multiplication employs additions and subtractions. Therefore, architecture of a multiplier for FPGAs and ASICs appears to be different. However, after dedicated carry logic has been incorporated into FPGAs, a ripple carry adder occupies half of the previous area and its speed has increased rapidly. This improvement has not been considered to re-establish the actual relation between the MM and LM architectures, and therefore the MM architecture has been overlooked. Summing up, in this exploration, the LM and MM architectures and their cost/speed relations in FPGAs are investigated.

A constant coefficient multiplier is usually implemented in a multiplierless fashion by using only shifts and adders from the binary representation (BR) of the multiplicand. For example, A multiplied by B= 14= 11102 can be

implemented as (A<<1)+(A<<2)+(A<<3), where '<<' denotes a shift to the left. It should be noted that the hardware requirements depend on the choice of a coefficient, i.e. the number of 1's in the binary representation of the coefficient should be as low as possible. Therefore several algorithms have been developed in order to reduce hardware by a proper choice of the multiplication coefficient (e.g. for FIR filters design). However, in this examination the assumption is made that the value of the coefficient is an input parameter to the design and therefore the coefficient value cannot be changed.

a. Convolution and correlation

Convolution can be used to compute the response of a linear system to an input signal. This linear system is defined by its impulse response. The output signal response is convolution of the input signal and the impulse response. Digital filtering is accomplished by determining a linear system´s impulse response that when convolved with the signal accomplishes the desired result (low-pass or high-pass filter).

The correlation algorithm is very similar mathematically to convolution, although, it is used for different purposes. It is usually used to identify the time delay at which two signals "line up", or are "most similar".

b. Weighted Partitioning

Graph clustering is a central problem in graph mining and network analysis. In tuitively, it aims to group the vertices into clusters with strong intra-cluster connectivity and minimal inter-cluster connections. A large number of graph clustering algorithms have been developed. However, these algorithms typically must load and access the entire graph (into main memory), and the computational cost of these graph clustering algorithms can also be very expensive. As the graph size becomes bigger and bigger (a graph with tens and hundreds of millions of vertices are become quite common), the existing approaches becomes prohibitively expensive for the massive graphs.

c. Distribution Arithmetic

In order to reduce hardware cost, all the bits of an input data can be processed in a parallel manner using Look Up tables. But in Distributed Arithmetic structure, bitwise operation is performed. The input data is processed bit by bit, first processing the least significant bit and then rest of the bits. There is substantial reduction in hardware as all bits have to pass through same architecture. In other words, if there is a N-bit parallel design, the DA method requires only (1/N)th of the hardware resources. As a result only one clock cycle is required for execution while about N cycles are required in serial execution. However, for the serial design the time-hardware product is smaller than the parallel design because the propagation delays are generally smaller as compared to the parallel structure.

III. PRIOR WORK

SR. NO.	TITLE	AUTHORS	YEAR	APPROACH
1	Weighted Partitioning for Fast Multiplierless Multiple-Constant Convolution Circuit	G. D. Licciardo, C. Cappetta, L. Di Benedetto and M. Vigliar,	2017	A new radix-3 partitioning method of natural numbers, derived by the weight partition theory, is employed to build a multiplierless circuit that is well suited for multimedia filtering applications
2	Low complexity and low power multiplierless FIR filter implementation,	X. Lou and W. Ye,	2017	For the implementation of multiplierless FIR filters, the product accumulation block (PAB) in the transposed direct form (TDF) structure has been ignored for a long time
3	Multiplierless two-stage comb structure with an improved magnitude characteristic	G. J. Dolecek and A. Fernandez-Vazquez,	2016	A simple multiplierless two-stage comb decimation filter. The filter exhibits an improved magnitude characteristic in both pass band and comb folding bands.
4	Efficient design and implementation of multiplierless FIR filter,	K. H. Dangra and G. S. Gawande,	2016	Introduced a novel design methodology for implementing Multiplierless FIR filter in hardware
5	Verilog implementation of fully pipelined and multiplierless 2D DCT/IDCT	G. R. Teja, R. S. Sruthi, K. S. Tomar, S. Sivanantham and K.	2015	The VLSI Implementation of a fully pipelined multiplier less architecture of 2D DCT/IDCT

	JPEG architecture	Sivasankaran		has been studied.
6	Design of high-speed multiplierless linear-phase FIR filters	W. B. Ye, X. Lou and Y. J. Yu,	2015	we propose a speed oriented optimization of linear phase FIR filters, where the length of critical path is used as the criteria in the discrete coefficient search
7	A New Class of Low Complexity Low-Pass Multiplierless Linear-Phase Special CIC FIR Filters,	D. N. Milić and V. D. Pavlović	2014	A new class of selective low-pass multiplierless linear-phase special CIC FIR filter functions of low complexity given in an explicit cascaded compact form. Several examples of the proposed filters are illustrated
8	Low-Power, High-Throughput, and Low-Area Adaptive FIR Filter Based on Distributed Arithmetic	S. Y. Park and P. K. Meher,	2013	This brief presents a novel pipelined architecture for low-power, high-throughput, and low-area implementation of adaptive filter based on distributed arithmetic (DA).

G. D. Licciardo, C. Cappetta, L. Di Benedetto and M. Vigliar [1] A new radix-3 partitioning method of natural numbers, derived by the weight partition theory, is employed to build a multiplierless circuit that is well suited for multimedia filtering applications. The partitioning method allows conveniently premultiplying 32-b floating-point filter coefficients with the smallest set of parts composing an unsigned integer input. In this way, similar to the distributed arithmetic, shifters and recoding circuitry, typical of other well-known multiplier circuits, are completely substituted with simplified floating-point adders. Compared to the existent literature, targeted to both field-programmable gate array and std_cell technology, the proposed solution achieves state-of-the-art performances in terms of elaboration velocity, achieving a critical path delay of about 2 ns both on a Xilinx Virtex 7 and with CMOS 90-nm std_cells.

X. Lou and W. Ye, [2] For the implementation of multiplierless FIR filters, the product accumulation block (PAB) in the transposed direct form (TDF) structure has been ignored for a long time. In this work, the hardware complexity and power consumption of the PAB is investigated. It is shown that the PAB contributes the majority of hardware complexity and power consumption in multiplierless FIR filter circuits. Implementation methods for the PAB are proposed to reduce the overall hardware complexity or power consumption of multiplierless FIR filters. Experimental results show that the overall hardware complexity and power consumption can be reduced significantly.

G. J. Dolecek and A. Fernandez-Vazquez, [3] This examination work presents a simple multiplierless two-stage comb decimation filter. The filter exhibits an improved magnitude characteristic in both pass band and comb folding bands. The attenuation in folding bands is increased by inserting additional zeros into comb folding bands, which are provided by two additional combs at the second stage. The comb pass band droop is decreased by cascading a simple compensator at low rate. The corresponding structure is simple and can be implemented either in nonrecursive or recursive form as CIC (Cascaded-Integrator-Comb) filter. The comparisons with some existing comb-based decimation filters demonstrate benefits of the proposed method.

K. H. Dangra and G. S. Gawande [4] FIR filter is a basic part used in Digital Signal Processing application because of its linear phase, stability, low cost and simple structure. The drawback of FIR digital filters is its high hardware complexity due to large number of multiplications. This research introduced a novel design methodology for implementing Multiplierless FIR filter in hardware. The proposed algorithm aiming at the reduction of hardware cost by minimizing the number of sign power of two terms (non zero terms) in filter coefficient. In this examination, implemented multiplierless FIR filter with and without optimized coefficients and their performances are compared in terms of speed, power, and area. It is observed that the power consumption for multiplierless FIR filter with unoptimized coefficient is 0.182W and the power consumption for multiplierless FIR filter with optimized coefficient is 0.176W. It also consumes fewer resources as compare to unoptimized coefficient multiplierless FIR filter.

G. R. Teja, R. S. Sruthi, K. S. Tomar, S. Sivanantham and K. Sivasankaran, [5] The concept of image compression is widely used in many fields like academics, industry and commerce for the transmission of data at higher speed and to allow the storage of large amount of data in less space. In this exploration work the VLSI Implementation of a fully pipelined multiplier less architecture of 2D DCT/IDCT has been studied. The compression and decompression is carried out with the help of two 1D-DCT calculations and a transpose buffer. The main objective is

to illustrate the improvement in the existing lossy compression design of JPEG by the introduction of pipelining and the introduction of BinDCT multiplier less architecture based on Loeffler's factorization. The design and implementation is carried out using verilog code.

W. B. Ye, X. Lou and Y. J. Yu [6] In the design of multiplierless FIR filters, researchers have made every effort to reduce the number of adders when coefficients multipliers are realized using adder-and-shift network to decrease the overall chip area. However, with the advance of IC technology, area becomes a less important issue than the speed. In this examination, this work reported a speed oriented optimization of linear phase FIR filters, where the length of critical path is used as the criteria in the discrete coefficient search. The length of critical path is measured as the number of cascaded full adders rather than the traditional adder depth. Compared to the area oriented algorithm, the proposed algorithm can generate the filters with much shorter critical path delay and meanwhile the area-delay product is also reduced. Gate level simulations of benchmark filters verify the above claim.

D. N. Milić and V. D. Pavlović,[7] This letter presents a new class of selective low-pass multiplierless linear-phase special CIC FIR filter functions of low complexity given in an explicit cascaded compact form. Several examples of the proposed filters are illustrated. They are compared in a fair way with conventional CIC FIR filters for the equal values of group delay and the same number of cascade sections. An example of the new class of the proposed seventh order filter with nine cascades is presented, where its stopband attenuation has high value of 143.82 dB, while corresponding conventional CIC FIR filter with the same filter order and the number of cascades has only 115.18 dB.

S. Y. Park and P. K. Meher [8] this brief presents a novel pipelined architecture for low-power, high-throughput, and low-area implementation of adaptive filter based on distributed arithmetic (DA). The throughput rate of the proposed design is significantly increased by parallel lookup table (LUT) update and concurrent implementation of filtering and weight-update operations. The conventional adder-based shift accumulation for DA-based inner-product computation is replaced by conditional signed carry-save accumulation in order to reduce the sampling period and area complexity. Reduction of power consumption is achieved in the proposed design by using a fast bit clock for carry-save accumulation but a much slower clock for all other operations. It involves the same number of multiplexors, smaller LUT, and nearly half the number of adders compared to the existing DA-based design. From synthesis results, it is found that the proposed design consumes 13% less power and 29% less area-delay product (ADP) over our previous DA-based adaptive filter in average for filter lengths N = 16 and 32. Compared to the best of other existing designs, our proposed architecture provides 9.5 times less power and 4.6 times less ADP.

IV. PROBLEM IDENTIFICATION

Technological developments have caused acceleration in the development of the area of DSP and GPUs. One of them is the devising of an efficient algorithm to calculate the Discrete Fourier Transform. The introduction of programmable digital signal processors (PDSPs) in the late 1970s was another major step in this field. These were able to perform “Multiplication and Accumulation” (MAC) in one clock cycle . It was an important improvement in the “Von Neumann” microprocessor based systems in those years. More improved functions such as memory bank, floating point multipliers, or zero overhead interface to ADC and DAC are included in the modern PDSPs. Digital Signal Processing finds major use in decoding, coding, image compression, audio and speech processing.

With design for low power gaining much importance for next generation technologies, various power reduction techniques are being proposed and adopted for reducing power in a VLSI circuit. Signal processing and communication blocks being the major power consuming blocks that perform high speed data processing activity, it is required to reduce power in these blocks by use of novel approaches. The problem statement in this research work adheres to this requirement. In this work, power reduction techniques that minimize power dissipation are analyzed, new techniques are examined to reduce power at all levels of abstraction (circuit level, sub-system level and architecture level).

V. CONCLUSION

This work presents an extensive survey of literature on weighted partitioning for fast multiplierless multiple-constant convolution circuit. By extending the analysis of prior work and providing new insight, in many cases multiplier less multiple constant multiplication algorithms are able to produce solutions in less run time with no more adders than the existing algorithms. In this algorithm, this examination has introduced aggressive pruning methods. Due to this, it empowers to solve larger problem sixes within a reasonable amount of time. Heuristics are needed for problems that are presently too large to exhaustively search in practice.

REFERENCES

[1]. G. D. Licciardo, C. Cappetta, L. Di Benedetto and M. Vigliar, "Weighted Partitioning for Fast Multiplierless Multiple-Constant Convolution Circuit," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 64, no. 1, pp. 66-70, Jan. 2017.

[2]. X. Lou and W. Ye, "Low complexity and low power multiplierless FIR filter implementation," 2017 IEEE 12th International Conference on ASIC (ASICON), Guiyang, 2017, pp. 596-599.
[3]. G. J. Dolecek and A. Fernandez-Vazquez, "Multiplierless two-stage comb structure with an improved magnitude characteristic," 2016 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), Jeju, 2016, pp. 607-610.
[4]. K. H. Dangra and G. S. Gawande, "Efficient design and implementation of multiplierless FIR filter," 2016 International Conference on Computing Communication Control and automation (ICCUBEA), Pune, 2016, pp. 1-5.
[5]. G. R. Teja, R. S. Sruthi, K. S. Tomar, S. Sivanantham and K. Sivasankaran, "Verilog implementation of fully pipelined and multiplierless 2D DCT/IDCT JPEG architecture," 2015 Online International Conference on Green Engineering and Technologies (IC-GET), Coimbatore, 2015, pp. 1-5.
[6]. W. B. Ye, X. Lou and Y. J. Yu, "Design of high-speed multiplierless linear-phase FIR filters," 2015 IEEE International Symposium on Circuits and Systems (ISCAS), Lisbon, 2015, pp. 2964-2967.
[7]. D. N. Milić and V. D. Pavlović, "A New Class of Low Complexity Low-Pass Multiplierless Linear-Phase Special CIC FIR Filters," in IEEE Signal Processing Letters, vol. 21, no. 12, pp. 1511-1515, Dec. 2014.
[8]. S. Y. Park and P. K. Meher, "Low-Power, High-Throughput, and Low-Area Adaptive FIR Filter Based on Distributed Arithmetic," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 60, no. 6, pp. 346-350, June 2013.
[9]. S. L. Chen, “VLSI implementation of an adaptive edge-enhanced image scalar for real-time multimedia applications,” IEEE Trans. Circuits Syst. Video Technol., vol. 23, no. 9, pp. 1510–1522, Sep. 2013.
[10]. F. C. Huang, S. Y. Huang, J. W. Ker, and Y. C. Chen, “High perfor- mance SIFT hardware accelerator for real-time image feature extraction,” IEEE Trans. Circuit Syst. Video Technol., vol. 22, no. 3, pp. 340–351, Mar. 2012.
[11]. G. D. Licciardo, A. D’Arienzo, and A. Rubino, “Stream processor for real- time inverse tone mapping of full-HD images,” IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 23, no. 11, pp. 2531–2539, Nov. 2015.
[12]. M. Vigliar and G. D. Licciardo, “Hardware coprocessor for stripe-based interest point detection,” US Patent 20 130 301 930, Nov. 14, 2013.
[13]. K. K. Parhi, VLSI Signal Processing Systems: Design and Implementation. New York, NY, USA: Wiley, 2007.