The Optimized data path ANN for Low power and Embedded applications.

S N Prasad [# 1]

#1.Research Scholar,
Jain University, Assistant Registrar(Evaluation), School of ECE,
REVA University, Bangalore,India
snprasad.aithal@gmail.com

S.Y.Kulkarni [#2]

#2.Principal Dirctor
REVA University,Bangalore,India
kulkarni_sy@reva.edu.in
sy_kul@yahoo.com

***Abstract* - This present work is aimed at the optimization of ANN (artificial neural network) for the low power & embedded applications. Due to rapid switching of the internal signals, power dissipation is very high in the modern VLSI systems. So the optimization is very much essential. This work explores the approaches to modify the existing building blocks of ANN in order to reduce the power (data path optimizations).by considering the 4:2 compressor architecture for the multiplier architecture of layered ANN. The design is modeled using Verilog HDL in the ASIC domain using the CMOS technological library of 65nm.The modified data path architecture consumes 15.91% of area and 26.09% of leakage power lesser when compared with existing architectures. This design provides the better speed up to 12.71%.**

Keywords: ANN, Multiplier, Compressor, Low power, VLSI, Verilog

I. INTRODUCTION

The important issue present in today's consumer electronics is the low power. Enhancing the processing performance and reduce the power consumption of the circuit designs are undoubtedly having the challenges in low power VLSI Design & other embedded applications..ANN is the system which contains small processing units called as neurons. These neurons perform the specific task in parallel .In ANN the neurons are arranged in the layered form, in such a way that the output of the neurons in a layer becomes the input to the neurons in the next. Each neuron contains N bit input data and weight.ANN composed of three layers named as input, hidden and output layers respectively. ANN supports parallelism methodology and dedicated hardware architectures are used to perform parallel operations. Hence it is easy to build the architecture for the ANN structures, which consists of simple similar operations in parallel manner. This enables them to have huge computational load [1].

The simple arithmetic operations and many special purpose architectures were developed using the regular ANN structure and in order to overcome the implementation problems it is matched to the integrated circuit technology by specifying the design and layouts of the VLSI circuits [3]. Due to the dependency on stored power supplies for the power in the portable electronic devices, power consumption plays an important role. Many of research organizations have focused on the development of low power implementations and methodologies.

When compared with the digital computers the BNN (Biological neural network) provides efficient performance for the tasks like vision, speech and image. From the research it is been observed that the visual system of a human does more image processing when it is compared with that of world's supply of super computers [4]. In contrast to the conventional processors ANN contains larger number of processing elements (PE's) and inturn in order to process and carry the information from one element to another element it uses adders, multipliers and memory caches. Hardware implementation of ANNs facilitates high functional capabilities to meet the desired constraints of real-world problems.

Advancement in VLSI technology enables enough flexibility in achieving the desired goals for a given application [5].

In the past implementations in order to reduce the power,the methods like variable supply voltage, clock gating, algorithmic optimizations were used. But due to the improvement in the scaling (technology scaling) one has to develop the approach specific to the particular application, data path optimization (for the power reduction). In the present work, the multiplier architecture has been chosen for the optimization, so that the modified architecture requires less amount of power. In the ealier implementations several low power multiplier architectures have been proposed in which power reduction is achieved by the method called structural

modification. In the architecture of the multiplier mentioned in the [6], power reduction is achieved by disabling the entire partial product row, if the multiplier bit corresponding to it is zero. Similarly for multiplier designs in [7] it disables the column bits of the partial product rows, if the multiplicand bit corresponding to it is zero for the power reduction. And for design of multiplier in [8] the Add and shift based multiplier design was used to lower the power consumption in the filter design. In the same way for the design mentioned in [9] co-efficient scaling and encoding techniques were adopted in order to reduce the power consumed by the design.

In the above mentioned work, power reduction methods for the multiplier design is done by the structural modifications as per the specific constraint. And moreover in this lower submicron technology, power consumption is more critical and the above architectures mentioned architectures saturate to achieve further power reduction. Hence one viable solution is to achieve the desired constraint through the optimizations of the critical and most copiously used data path architectural components. The Ravi et al in [10] proposed the data aware Brent Kung adder in stage of carry propagation of Dadda multiplier design, in order to illustrate the importance of the critical components of the design. In the submicron technology node leakage power is the major constraint, even though this architecture has lower power consumption but limits when leakage power is the major constraint.

In this data path architecture, optimizations are provided for multiplier architecture of the 3 layered ANN design. Here the compressor architecture present in the multipliers are optimized to address the leakage power of the design.Carefull attention is paid for the other parameters of the design, while improving the leakage power of the compressor architecture. Suitable measures were taken in such a manner that the proposed design doesn't affects the other parameters of the design like area and delay. Next sections of this work includes an architectural aspects and its importance w.r.t 3- layered ANN , Synthesis results, evolution and discussions of multiplier and ANN architecture,and finally the conclusion of the proposed design.

II. ANN ARCHITECTURE

This section deals about complete ANN architecture and its working along with its data path. Figure 1 shows the 3 layer ANN architecture; in which the number of multipliers per neuron will be equal to number of connections to this neuron and the number of adders will be equal to number of connections to previous layer minus one [11].

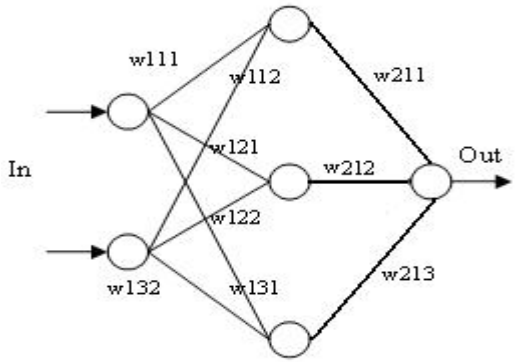

Figure 1: A 3-layer ANN

ANN is the 3 layered structures composed of input layer, hidden layer and the output layer. And here each neutron consists of N bit input and weight, and the output of the neuron has two stages multiplication and accumulation respectively. The number of bits required by the multiplication is 2N bits, but the next stage neutron has only N bit input so for this reason multiplication is truncated from 2N to N bits. Therefore the activation function for neuron is chosen to be truncating/rounding unit. This type of truncation results in the positive errors (bugs) in the range of '0' to '2^{-N-1}'. To minimize this error, the solution is provided in [12], in all applications accuracy plays critical role, but several applications can tolerate the inaccuracies only to certain level in order to minimize the computational errors. For example let us consider image processing applications, in which if the proposed image having the less accuracy provides same information when compared with that of more accurate design. Hence additional complexity is involved in accurate design can be trade off. This truncation method has following advantages such as it reduces the complexity involved in the computations and inturn the area, power and delay parameters.

ANN architecture is composed of multipliers and adders. Here the Wallace tree multiplier and ripple carry adder are used in the design. Here the multiplier consists of 3 stages named as partial product generation; partial product reduction and carry propagate addition. Out of these 3 stages, the second stage that is partial product reduction stage is the critical stage because it involves huge amount of computations, inturn this particular stage only decides the performance and power consumption of the design. The number of steps in partial product reduction stage can be reduced by using the compressors through the multi point additions. These compressors

perform the computations parallel manner; by this the critical path of the multiplier can be reduced. Let us consider the 8 bit multiplier, it requires 14 4:2 compressors and the number of compressors required for the multipliers increases nonlinearly with the increase in the bit width of the multiplier.

There exist several ways to implement the Compressors ,such as 3:2, 4:2, 5:2, and 7:2 In the past implementations ,some of the architectures have been developed in order to improve the performance and to reduce the power consumption. The Figure 2 represents the dot diagram of 4:2 compressor used in the Wallace tree multiplier for the multi input column addition.

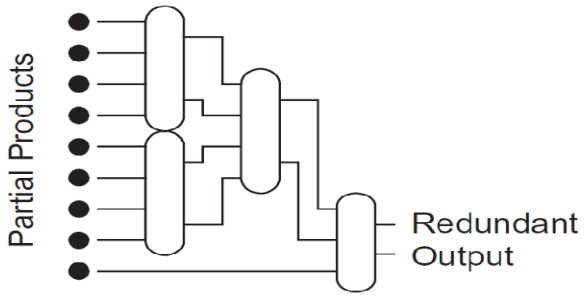

Figure 2: 4:2 compressor dot diagram for multiplier [11]

The compressor 4:2 as name itself indicates it takes 4 weighted inputs and two outputs named as sum and carry to the same and next column respectively. This compressor consists of inputs and outputs along a symbolic view of the 4:2 compressor architecture is shown in Figure 3. {A, B, C, D} and "CIX" are the four primary and one intermediate carry inputs. "S" is the output of the same column, "CO" is the carry out of the final stage to next column and "COX" is the intermediate carry-out to the next columns. 4:2 compressors can also be described as the cascade of Full-adders/carry save adders/3:2 compressors [14] as shown in Figure 4.

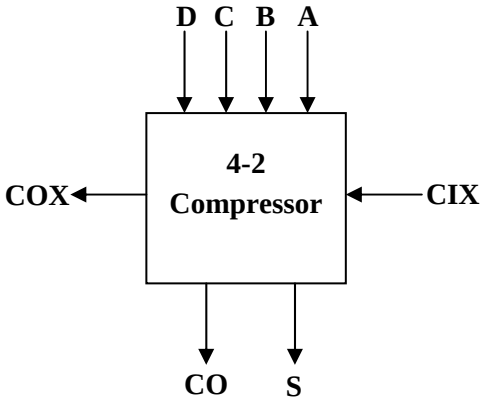

Figure 3: 4-2 compressor Symbol

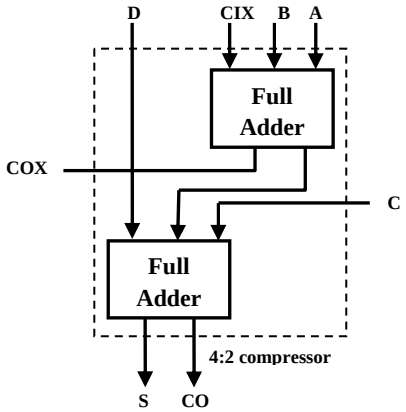

Figure 4: CSA/ Full adders based 4-2 compressor [13]

Typically or conventionally used Full adder architecture is shown in Figure 5. Conventional Full adder shown in Figure 5, consists of two XOR gates, two AND gates and one OR gate. As the number of cells are more in this architecture, interconnects between the gates will be more and results in the larger interconnect delays. These interconnects may also lead to glitches and dissipates more power. Since more number of smaller fan-in gates, the area required will be more and leads to higher power consumption and larger delays. Hence to mitigate such recurring effects, full adder cell with larger fan-in was proposed in this brief. Figure 6 shows the proposed full adder architecture.

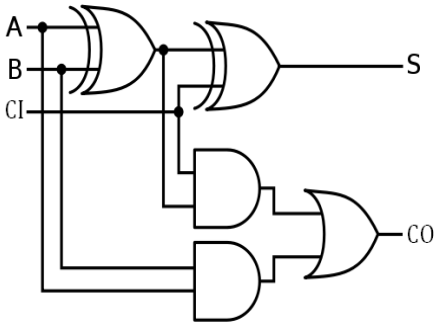

Figure 5: Conventional Full Adder architecture [12]

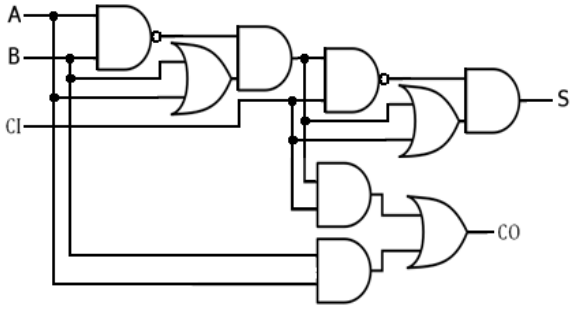

Figure 6: Proposed Full Adder architecture

The proposed full adder architecture contains large fan-in gates to achieve its functionality. Larger fan-in gates allow merging of smaller-in gates to form a single complex cell. Here two AND gates and one OR gate are merged to form a single complex AND-OR logic cell (AO22). Apart from this an alternate XOR-gate functionally equivalent low power gate level architecture was also proposed to lower the power consumption of the cell. Leakage power reduction was given prime importance in developing the gate level architectures.

Some of the advantages of the proposed architectures are as follows.

Use of Larger fan-in gates

- reduces the gate counts (less area)
- Increases the transistor stack which helps in the ON resistance between supply rails and reduces the leakage power consumption
- Reduced gate count reduces interconnects and associated interconnect delays and glitches. This results in reduced delay and dynamic power consumption

Optimizations are applied in the following parts of the ANN architecture

- In compressor of multiplier
- For full adders other than the compressor architecture in multiplier
- For full adders in the adder stage of ANN architecture (since RCA is the cascade of Full-adders)

III. RESULTS & OBSERVATIONS

The conventional and proposed 3 layer ANN architecture design is modeled by using the Verilog HDL language. The functionality of conventional and proposed designs are examined by using the model sim simulator's waveform editor. The synthesis of the design is carried out with the ASIC methodology, using the RTL compiler of the cadence by targeting the 65nm CMOS technology library [15-16]. This proposed architecture is applied to all the possible components of the ANN architecture and with this an observation is made on the importance of the data path of the design. The results obtained from the proposed method execution are compared with the conventional design results at the various parts of the ANN architecture, and these results are tabulated in the table's I-IV.

TABLE I: Results of the ANN with proposed compressor architecture

Parameter	ANN with existing architectures	ANN with Proposed compressor architecture	%change
area (square microns)	9849.96	9073.8	7.87
delay (nano seconds)	7.54	7.034	6.80
Dp (micro watt)	992.75	953.061	3.99
Lp (micro watt)	84.26	72.827	13.57
Tp (micro watt)	1077.02	1025.888	4.74

Note: Dp → dynamic power; Lp → leakage power; Tp →total power

TABLE I represent the results of ANN's conventional one and proposed one. From the table it has been observed that proposed architecture results in better in terms of all parameters when compared with conventional design. This proposed architecture is built by paying more attention to the reduction of leakage power. The exact outplay of proposed design is given in Table I. When compared with the existing architecture the proposed architecture has reduced 13.57% of the leakage power, 7.87% less area, and 6.8% speeder (faster) than the conventional architecture of ANN. The multiplier required for this proposed architecture is only the 8 bit multiplier which requires 14, 4:2 compressors were optimized in single multiplier. For high bit width multipliers ,the number of compressors required will be more ,hence optimizations requireires d is also more.ANN architecture mentioned in figure 1 requires totally 11 multipliers, hence totally 154 compressors architectures has to be optimized .

TABLE II represents the results of ANN architecture, when proposed architecture is applied to the Full adders and also when proposed architecture is applied to the addition part of the ANN architecture. When proposed architecture is applied to all full adders of the multiplier design of the ANN, the obtained results are tabulated in TABLE III.

Here the impact of optimization on design is less because, the addition part of ANN architecture is less than the number of compressors architectures present in the entire ANN architecture. In this proposed concept the importance is given to the reduction of leakage power and this proposed technique is applicable for any level at any hierarchy levels of the design cycle.

TABLE II: Results of the proposed architecture in addition part of ANN architecture

Parameter	ANN with existing architectures	ANN with Proposed Adder architecture	%change
area (square microns)	9849.96	9657.36	1.95
delay (nano seconds)	7.54	7.44	1.41
Dp (micro watt)	992.75	981.99	1.08
Lp (micro watt)	84.26	81.58	3.18
Tp (micro watt)	1077.02	1063.58	1.24

Note: Dp → dynamic power; Lp → leakage power; Tp →total power

TABLE III: Results of the proposed architecture in partial product reduction and carry propagate addition stage of multiplier of ANN architecture (conventional compressor architecture)

Parameter	ANN with existing architectures	ANN with Proposed full adder in multiplier architecture	%change
area (square microns)	9849.96	9248.04	6.11
delay (nano seconds)	7.54	7.23	4.19
Dp (micro watt)	992.75	955.10	3.79
Lp (micro watt)	84.26	76.40	9.33
Tp (micro watt)	1077.02	1031.50	4.22

Note: Dp → dynamic power; Lp → leakage power; Tp →total power

The TABLE IV represents the results obtained when proposed concept was applied to the parts which are mentioned in the Table III.

TABLE IV: Results of the proposed architecture in entire ANN architecture

Parameter	ANN with existing architectures	ANN with Proposed architecture	%change
area (square microns)	9849.96	8279.28	15.94
delay (nano seconds)	7.54	6.58	12.71
Dp (micro watt)	992.75	904.60	8.87
Lp (micro watt)	84.26	62.27	26.09
Tp (micro watt)	1077.02	966.88	10.22

Note: Dp → dynamic power; Lp → leakage power; Tp →total power

The data path optimizations have impact significantly on the all parameters of the design; These results are tabulated in Table IV. When compared to the conventional design's leakage power the proposed design has 26.06% of leakage power, area is reduced by amount of 15.94% when compared with the previous designs. The processing speed of the proposed design is 12.71%more than the conventional design.

The data path architectural optimizations results are obtained from all above tables are found to be unique. By this feature ,it can be applied to any level of the hierarchy of the design cycle for any bit width according to the application requirements,(Table III) and addition stage (Table II) of ANN architecture).

IV. CONCLUSION

In this work, the 3- layered ANN architecture is illustrated in the ASIC domain for the low power application. Here mainly the data path optimization techniques are proposed on ANN architecture for the lower power application. Especially for battery powered devices, these runs over lower technological node cells; in which leakage power is the major concern. The architecture obtained from the data path optimizations (proposed design) architecture has reduced leakage power of 26.09% and 15.94% of area. The speed of the proposed architecture is 12.17% more than the existing design without any optimizations. From the analysis of the results obtained at the different blocks of the ANN, it is clear that optimizations made are unique and these optimizations can be applied to any level of the hierarchy in the design cycle for any bit width of the design as per application requirement.

V. ACKNOWLEDGEMENT

I sincerely thank JAIN University for providing me an opportunity to carry out my research by providing the necessary guidance,encouragement and fullest support in all my research endeavours.I would like to pay my gratitude towards my affiliation REVA University for providing me all the necessary management and moral support in carrying out the work.I would fail in my duties if I don't mention my mentor,research guide Dr.S.Y.Kulkarni,the co-author of this article, who supported me in all respects.

VI. REFERENCES

[1] Michel Verleysen, "VLSI implementations of artificial neural networks" UCL, Dec 2000. http://perso.uclouvain.be/michel.verleysen/papers/agregation_ancillary3.pdf
[2] Asanovic, Nelson Morgan1 Krste, Brian Kingsbury, and John Wawrzynek. "Developments in Digital VLSI Design for Artificial Neural Networks." University of California at Berkeley, Berkeley, California (1990).
[3] Boser, Bernhard E., et al. "Hardware requirements for neural network pattern classifiers." IEEE Micro 12.1 (1992): 32-40.
[4] C. A. Mead, Ismail M, "Analog VLSI Implementations of Neural Systems", Reading, MA: Addison-Wesley, 1989
[5] Fang, Xuefeng "Small area, low power, mixed-mode circuits for hybrid neural network applications" Diss. Ohio University, 1994.
[6] Ohban, J., Moshnyaga, V.G., and Inoue, K.: "Multiplier energy reduction through bypassing of partial products". Proc. Asia-Pacific Conf. on Circuits and Systems, 2002, Vol. 2, pp. 13–17
[7] Wen, Ming-Chen, Sying-Jyan Wang, and Yen-Nan Lin. "Low power parallel multiplier with column bypassing." Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on. IEEE, 2005.
[8] Rashidi, B.; Pourormazd, M., "Design and implementation of low power digital FIR filter based on low power multipliers and adders on xilinx FPGA," Electronics Computer Technology (ICECT), 2011 3rd International Conference on , vol.2, no., pp.18,22, 8-10 April 2011.
[9] Sangjin Hong; Suhwan Kim; Papaefthymiou, M.C.; Stark, W.E., "Low power parallel multiplier design for DSP applications through coefficient optimization,"ASIC/SOC Conference, 1999. Proceedings. Twelfth Annual IEEE International , vol., no., pp.286,290, 1999.
[10] Ravi, S., Nair, G. S., Narayan, R., & Kittur, H. M.. "Low Power and Efficient Dadda Multiplier", Research Journal of Applied Sciences, Engineering and Technology 9(1): 53-57, 2015
[11] Sahin, Suhap, Yasar Becerikli, and Suleyman Yazici. "Neural network implementation in hardware using FPGAs." Neural Information Processing. Springer Berlin Heidelberg, 2006.
[12] Dhafer r. Zaghar, "Reduction of the error in the hardware neural network", Al-khwarizmi Engineering Journal ,Vol.3, No. 1 PP, 80-41 (2007)
[13] Neil H Weste and David M Harris, "CMOS VLSI Design- A Circuits & System Perspective", Pearson Education, 2008.
[14] Aliparast, Peiman, Ziaadin D. Koozehkanani, and Farhad Nazari. "An Ultra High Speed Digital 4-2 Compressor in 65-nm CMOS." International Journal of Computer Theory & Engineering 5.4 (2013).
[15] Mentor Graphics Corporation, ModelSim SE Tutorial, 2008. http://www.mentor.com
[16] Cadence Design systems, Quick Reference for Encounter RTL Compiler, 2006. http:// www.cadence.com