International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 18 (2015) pp 38785-38789

Low Power Datapath Architecture for ANN

S N Prasad
Research Scholar, Jain University, Associate Professor, Dept. of ECE, Reva Institute of Technology & Management
Reva University, Bangalore snprasad.aithal@gmail.com

S Y Kulkarni
Principal Director-Academics, Reva University,Bangalore kulkarni_sy@reva.edu.in sy_kul@yahoo.com

Abstract
Low power gate level datapath optimizations are presented for Artificial Neural Network (ANN) architecture to address the low power ANN applications in the field of science and engineering. Efficient 4:2 compressor architecture is proposed for the multiplier architecture of ANN layered structure. Proposed datapath architectural optimizations are illustrated in the 2-3-1 tree layer artificial neural network (ANN). Verilog HDL was used to model the design in ASIC domain with 65nm technological CMOS library. The proposed concept has resulted with 12.71 % better speed, 15.94 % less area and 26.09 % less leakage power consumption.

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

Introduction

ANN's are recognized as the powerful technology in solving many real-time applications because of its massive parallelism. Dedicated hardware architectures to perform parallel operations make it perfectly suitable for VLSI implementations. It is simpler to build the architectures for the ANN structure; consisting of simple similar operations in parallel. This enables them to have heavy computational load [1].

Simple similar operations and heavy computational loads have made them popular with VLSI implementations and at the same time many interesting problems were generated by such implementations. Integration of large number of cells, optimal resource utilization, connectivity, information transmission and computational power requirements are some of the problems to be named. Digital implementations were approached to mitigate and provide viable solutions to such problems; which led to the development of several design methodologies for digital ANN [2]. With its regular structure and simple arithmetic operations many special purpose architectures were developed and matched to the integrated circuit technology to overcome the implementation problems by simplifying the design and layouts of VLSI circuits [3]. But processing speed of the ANN architectures is not the only constraint to be considered in lower technological node, however power consumption also plays a major role in portable electronics because of the dependency on stored power supplies. Hence many research organizations have focussed on the development of low power implementations and methodologies.

The biological neural network provides strong performance in tasks such as vision, speech, image and etc than the digital computer. It is observed that the visual system of a human does more image processing than the world's supply of super-computers [4]. Unlike the conventional processors, ANN contains large collection of processing elements (PEs), which inturn have adders, multipliers and memory cache to process and carry the information from one element to another. 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].

Several approaches like variable supply voltage, clock gating and some algorithmic optimizations were made in the past to mitigate the large power consumption. But due to scaling of technology, approaches have been saturated and there is a need to develop the architecture's specific to application by adopting the optimizations/constraint aware datapath architectures at the lowest hierarchical level in the design cycle. Hence in this brief multiplier architecture is optimized to lower the power consumption and make it suitable for compute intense applications where power consumption is a critical constraint. Several low power multiplier architectures were developed in the past. Power reduction can be achieved in multiplier through the structural modification. A multiplier design in [6] has the provision of disabling the entire partial product row if the corresponding multiplier bit is zero. Similar to the concept of [6], multiplier design of [7] disables the column bits of the partial product rows if the corresponding multiplicand is zero. Add and shift based multiplier architecture was used for the filter design in [8], to lower the power consumption. Minimum bit clustering and partitioning of the multiplier architecture was achieved by co-efficient scaling and encoding to lower the power consumption [9].

The above multiplier architecture's reduce the power consumption of the design through structural some modifications and optimizations, but doesn't have the control or flexibility over the datapath to design 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 datapath architectural components. In this regard, Ravi et al in [10] has proposed the data aware Brent Kung adder in the carry propagate addition stage of the Dadda multiplier to illustrate the importance of critical datapath components in the design. This architecture even though has reduced the power consumption but limits

when leakage power is of major constraint. Since leakage power is dominant in this sub-micron technological node.

Hence in this brief, datapath architectural optimizations are provided to the most copiously used and critical components of the multiplier architecture and illustrated its importance in the 3 layered ANN architecture. Here the concept has been applied to multiplier architecture; where compressor architectures have been optimized to address the leakage power of the design. While improving the leakage power of the compressor architecture other design parameters were also considered, so that the proposed architecture doesn't affects other parameters (area and delay) of the design.

Other sections of the paper are organized as mentioned below. Section II gives the architectural aspects and its importance with respect to the 3-layered ANN architecture. Section III provides the evaluation and discussion over the synthesis results of the multiplier and ANN architectures. Conclusion is provided in section IV and references are given in the last section.

ANN Architecture

This section represents the ANN's architecture, working and its datapath. Fully parallel feed-forward ANN architecture is illustrated to review the importance of datapath architectures. Figure 1 shows the 3 layer ANN architecture; where 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].

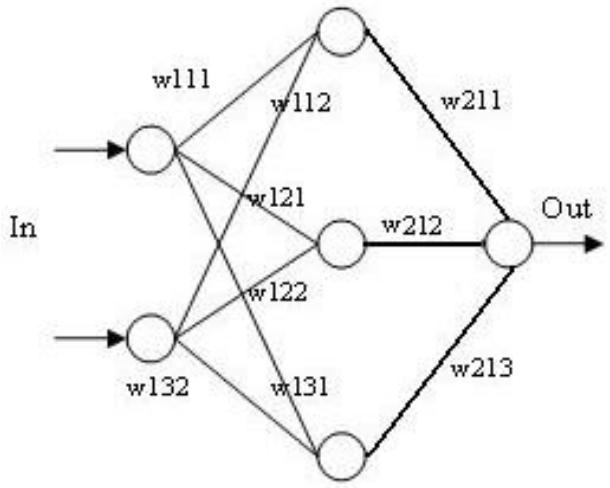

Figure 1: A 3-layer ANN

This structure consists of 3-layers, input, hidden and output layer. Each neuron consists of 'N' bit input data and 'N' bit input weight. The output of the neuron has two stages – multiplication and accumulation. Multiplication requires '2N' bits, but the input of the next stage neuron has 'N' bit input, therefore the multiplication has to be truncated to 'N' bit. Thus the transfer or activation function for the neuron is chosen to be the truncating/rounding unit. Such truncation introduces positive error between the range '0' to '2^{-N-1}'. However solutions are provided in [12] to minimize the error. Even though accuracy plays a critical role, several applications can tolerate inaccuracies to certain level so that their computational efforts are reduced. For example in image processing applications, if an processed image with less-accuracy provides an same information as that of the more-accurate design, then the additional complexity involved in accurate arithmetic design can be trade-off. Advantages of such truncation, reduces the complexity involved in the computations and its cumulative area, delay and power parameters.

As mentioned previously, the ANN architecture contains multipliers and adders. Here a typical Wallace tree based multiplier is utilized. And to implement the addition, a ripple carry adder (RCA) was incorporated. In multiplier, there are three stages namely partial product generation, partial product reduction and carry propagate addition. Partial product reduction is the critical stage as it involves large amount of computations and it decides the performance and power consumption of the design. Compressors are used to reduce the number of steps in the reduction stage through multi-input additions. And these compressors are computed parallely which help in reducing the critical path of the multiplier. A typical 8-bit multiplier contains 14 4:2 compressor components and these numbers increase non-linearly with the increase multipliers bit-width.

Compressors can be implemented in several ways (3:2, 4:2, 5:2, and 7:2) and various architectures were developed in the past to improve the power and performance of the multiplier architecture. Figure 2 shows the typical compressor dot diagram used in the tree multipliers for multi-input column addition.

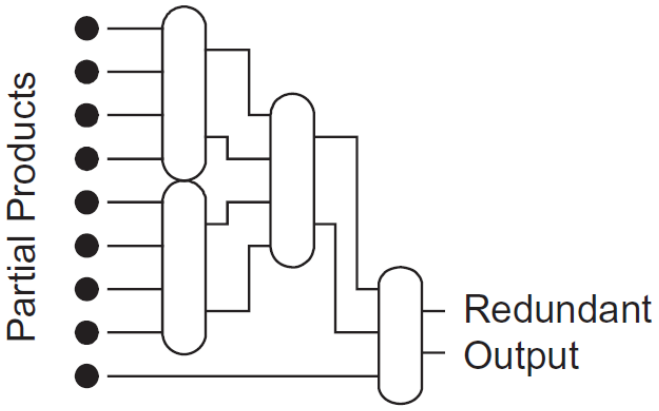

Figure 2: Dot diagram of 4-2 compressor in multiplier [13]

This 4:2 compressor takes equal weighted 4-inputs and generates two outputs – sum and carry to the same and next column respectively. Apart from these inputs-outputs it also considers an intermediate carry from previous column and generates a similar intermediate carry to the next column. Intermediate carries are popularly called as horizontal carry and final stage carry-out is called as the vertical carry.

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.

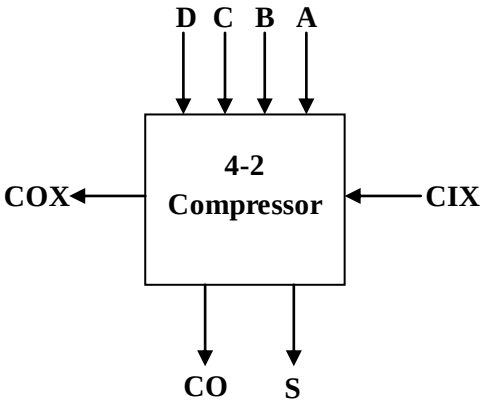

Figure 3: 4-2 compressor Symbol

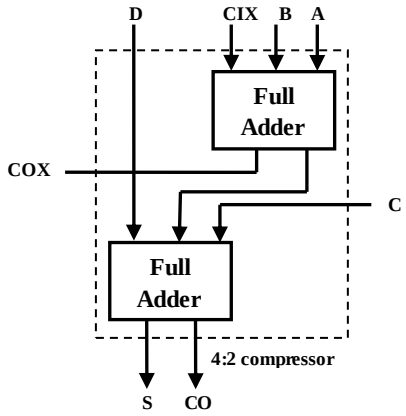

Figure 4: CSA/ Full adders based 4-2 compressor [14]

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's was proposed in this brief. Figure 6 shows the proposed full adder architecture.

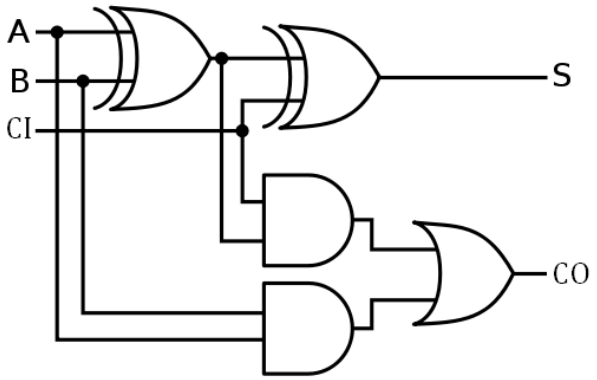

Figure 5: Conventional Full Adder architecture [13]

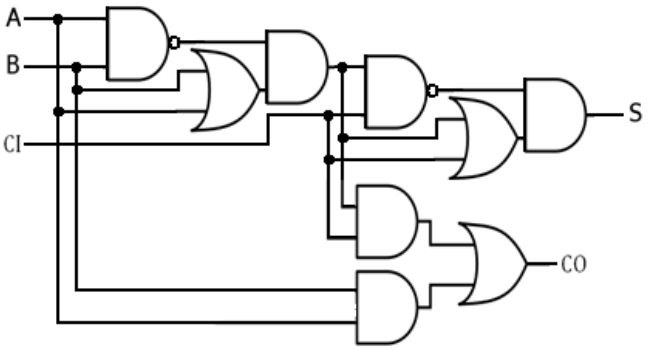

Figure 6: Proposed Full Adder architecture

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)

Results & Observations

3-layered ANN architecture with conventional and proposed architectures was modeled using Verilog HDL and their functionality was verified using model-sim simulator's waveform editor. Designs were synthesized with standard ASIC methodology using Cadence RTL compiler by targeting the CMOS 65nm technological library [15-17]. Proposed architecture was applied to all possible parts of the ANN architecture and the importance of datapath architectural optimizations are observed. The results of the conventional and proposed architectures (at various parts of ANN architecture) are tabulated in Tables 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 gives the results of the ANN with conventional and proposed compressor architectures. It shows that the proposed architecture has bettered the conventional architectures in all the design parameters. As mentioned in the previous that the proposed architecture was built with prime consideration of leakage power reduction and the Table I displays the exact outplay of the proposed architecture. It has reduced 13.57 % of leakage power than the counterpart existing architecture. It has also utilized 7.87% less area and process 6.8 % faster than the conventional ANN architecture. The multiplier required in this set-up was of only 8-bit and hence only 14 compressor cells were optimized in single multiplier. For larger bit-width multipliers the number of compressors will be more and the optimizations will also be more. Totally 11 multipliers are present in the ANN architecture of Figure 1 and hence totally 154 compressor architectures were optimized.

The proposed concept was applied to the full adders of the ANN architecture also and Table II displays the results when proposed architecture was applied to addition part of ANN architecture. Table III gives the results when proposed architecture was applied to all full adders of multiplier architecture in ANN architecture.

Here the optimization impact is less as the number of full adders in the addition part of ANN architecture is less than the number of compressor architectures in the entire ANN architecture. But the proposed concept of leakage power reduction importance hold good (higher than other parameters) even at this level and this proves that the proposed technique behave similarly at any hierarchical level 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

Similarly the proposed concept was applied to all above mentioned parts of the ANN architecture and their results are tabulated in Table IV.

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 results of the Table IV show that the datapath optimizations impacts significantly for all design parameters. The proposed architecture has reduced 26.09% of leakage power with 15.94 % less area and provided 12.71 % more processing speed to the architecture.
Results from all the Tables suggest that the datapath architectural optimizations are unique. Hence can be applied to any hierarchical level in the design cycle and for any bit width (for example in RCA at carry propagate addition stage in multiplier (Table III) and addition stage (Table II) of ANN architecture).

Conclusion

Low power 3-layered ANN architecture was illustrated in the ASIC domain in this paper. Low leakage power aware datapath architectural optimizations were proposed for low power ANN applications especially for battery powered devices running with lower technological node cells where leakage power is of prime concern. The proposed architecture has reduced 26.09% of leakage power with 15.94 % less area and provided 12.71% more processing speed to the architecture. Analysis of the proposed architectures results at different blocks of ANN architecture suggests that the datapath architectural optimizations are unique which can be applied to any hierarchical level in the design cycle and for any bit width.

References

[1] Michel Verleysen, "VLSI implementations of artificial neural networks" *UCL*, Dec 2000. http://perso.uclouvain.be/michel.verleysen/papers/aggregation_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] ChandraMohan U, "High Speed Squarer", *Proceedings of the 8th VLSI Design and Test Workshops, VDAT*, August 2004.

[16] Mentor Graphics Corporation, ModelSim SE Tutorial, 2008. http://www.mentor.com

[17] Cadence Design systems, Quick Reference for Encounter RTL Compiler, 2006. http:// www.cadence.com