Analyzing the Performance of 7nm FinFET Based Logic Circuit for the Signal Processing in Neural Network

Rajeev Kumar Pandey[1] and Sanjeev Kumar Pandey[2]
[1]EECS International Graduate Program, National Chiao Tung University Hsinchu, 300, Taiwan
[2]Department of Electrical Engineering, Indian Institute of Technology, Delhi, India
[1]rajeev.eed06g@nctu.edu.tw

2020 IEEE Recent Advances in Intelligent Computational Systems (RAICS) | 978-1-7281-9052-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/RAICS51191.2020.9332500

***Abstract*—This study aims to report the fan-out-four (FO4) delay and energy consumption associated with the various circuit used in the implementation of the neural network signal processing using the 7nm node. This study characterizes the behavior of the 7nm-FinFET to design the various logic circuits and then report the FO4 delay and the energy consumption associated with the basic gates, 16-bit Brent-Kung adder, Flip-flop, 16-bit array pipeline multiplier (clocked at 4GHz). The successful design of the 6T Static Random-Access memory (SRAM) and the clocked regenerative latch-based comparator (clocked at 4GHz) shows that the high-speed memory and mixed-signal circuit design is possible with 7nm FinFET. For the simulation, 7nm FinFET predictive Technology model (PTM) for bulk has been used with the BSIMCMG model. Device characterization shows that the drain current is proportional to the number of fins (NFIN). During the simulation, it is observed that the threshold voltage (0.3V) and the drain current of the N/P FinFET are almost equal. The measured leakage current is 3.5pA. The comparison of the simulation results with 45nm shows that the 7nm FinFET based logic design is highly efficient in terms of FO4 delay, power, and energy. Despite the reduced supply voltage in the 7nm node, the noise margin of the SRAM is still comparable to 45nm.**

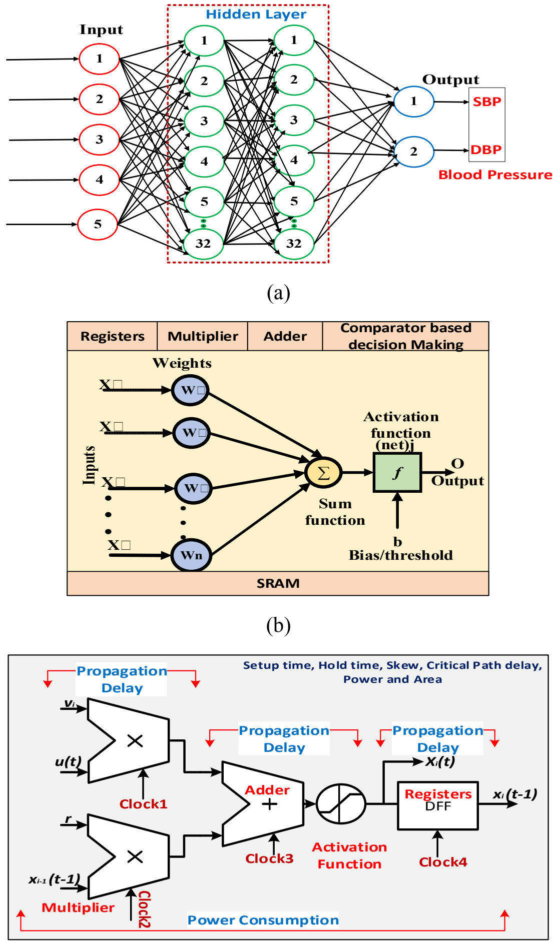
Keywords*—*7nm, FinFET, Regenerative Latch, SRAM, Adder, Multiplier, comparator, Flip Flops.

I. Introduction

According to the literature, the rapid growth of the Internet of Things (IoT) and the artificial neural network-based system, shows that the market of mobile computing would exceed 1 trillion dollars by 2022. Artificial neural networks (ANNs) are the computing system which is modelled to sense and process the information similar to the brain. An artificial neural network consists of several neurons units that communicate information to each other through a large number of weighted connections in the network. One of the examples of the simplified neural network (NN) structure is shown in Fig.1. NN incorporates the input layer, hidden layer, and output layer. Each node of the NN is called a neuron [1]. The fundamental equation of single neuron output (O) is,

$$O = f(w * x + b) \tag{1}$$

where ‘w’ is the weights, and ‘b’ is the activation function. The structure of the neuron is shown in Fig 1 (b). Fig 1(c) shows the architecture of discrete time neurons, which incorporates an adder, a multiplier, a Flip-Flop, a Static Random-Access memory (SRAM) memory, and a comparator. The output of the neurons is stored in the registers/SRAM so it can be used by another interfacing neuron in the next time step.



(a)

(b)

(c)

Fig.1. (a) Simplified Architecture of the neural structure (b) Basic components of a neural network. (c) Logical implementation of the NN.

It is well known that the modern NN architectures need a fine-grained and fast system for signal processing with minimum energy consumption. According to Fig. 1(c), to ensure the performance of the system, the prior knowledge of the highest operating frequency, propagation delay, and power consumption before the synthesis helps to design an efficient and accurate digital system. In general, the basic timing specification are setup time margin, hold time margin, critical path delay, skew and jitter. Furthermore, the low power implementation of the NN on the chip using advanced FinFET nodes (e.g. 22nm, 16nm, 12nm, 7nm, 5nm, and 3nm) are limited due to hard layout constraints, active power consumption and the leakage currents. Therefore, the prior knowledge of the timing specification and power consumption of each interfacing block ensures the proper functionality of the NN system. By considering all these aforesaid facts and challenges, in this study, 7nm technology will be explored to report the fan-out-four (FO4) delay and energy consumption associated with basic building blocks of the NN. Therefore, this study is divided into four sections. Section II presents the electrical characterization of the 7nm N-FinFET and P-FinFET. Section III will present the simulation results of various logic styles associated with NN. Finally, Section IV concludes this study.

II. FINFET CHARACTERIZATION AND SIMULATION SETUP

A. 7nm FinFET Characterization

The diagram of the FinFET is shown by Fig.2 [2]. FinFET is a three-terminal device that incorporates gate-drain and source as the three terminals. For the spice simulation, 7nm FinFET predictive Technology model (PTM) card version 105.03 for bulk is using. To invoke the BSIMCMG model from the HSPICE "MODEL NCH NMOS LEVEL=72, VERSION=105.03, BULKMOD=1" must be added in the model card. Here the term BULKMOD=1 means BSIMCMG Bulk model will invoked for the simulation [2][3][4]. If BULKMOD=0, this means that the BSIMCMG SOI model will be invoked for the simulation. The specified supply voltage for the 7nm FinFET is 0.7V. N-FinFET or P-FinFET can be used by "M1 drain gate-source NCH L=7n TFIN=1n NFIN=1". Point to be noted that the effective width of the device is decided by the NFIN therefore with an increase in the number of the fin (NFIN), the drain current increases proportionally**.** Note that the NFIN and fingers are only design parameters that are used to implement the various logic styles.

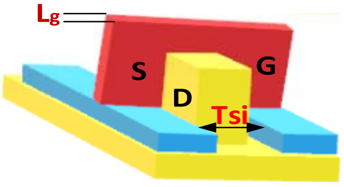
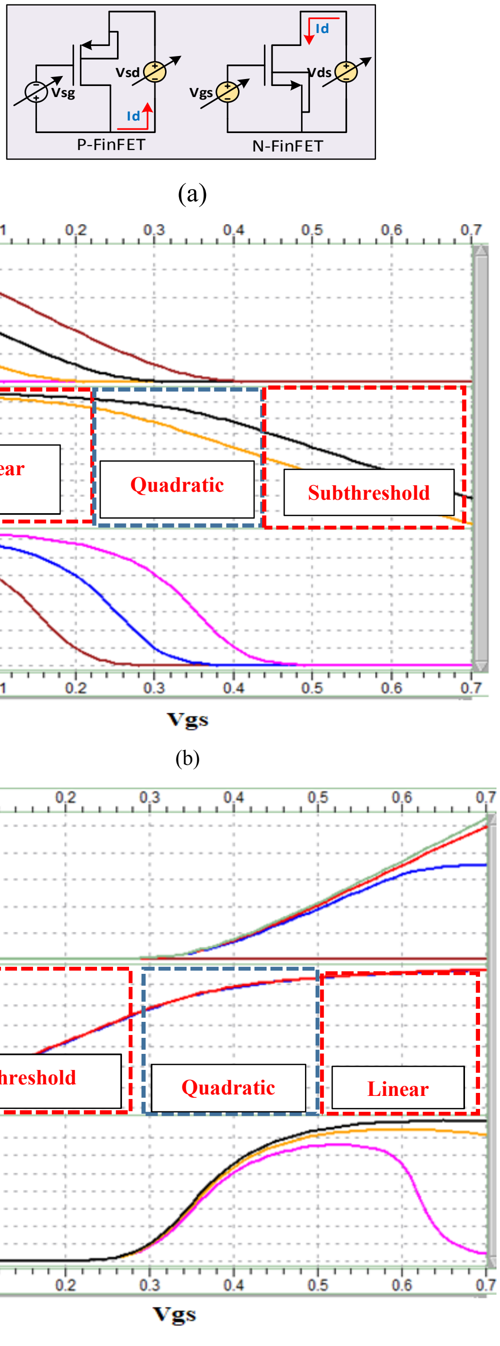


Fig. 2. FinFET Diagram

The simulation is performed by the keeping temperature 25 ℃, L=7nm, and NFIN=1. The test bench for the simulation is shown in Fig 3 (a). The drain current (ID) vs gate to source voltage (Vgs), log$_{10}$(ID) vs Vgs, and deriv (ID) vs Vgs for the P-FinFET has shown in Fig.3 (b). The results show that the threshold voltage for the P-FinFET is 0.3V. The plot log$_{10}$(ID) vs Vgs shows the subthreshold current and the boundary between the subthreshold and near-threshold or quadratic region. The derivative curve describes the boundary between the quadratic and linear region. Similarly, the ID vs Vgs, log$_{10}$(ID) vs Vgs, and derive (ID) vs Vgs for the N-FinFET are shown in Fig.3 (c). Also, the threshold voltage is nearly equal to around 0.3V. The above result shows that the drain current of the N-FinFET and P-FinFET are almost equivalent. Therefore, there is no need to follow the rule like P- FinFET size must be equal to 2- 3 times to N-FinFET for the same drain current as in CMOS technology. The ID vs V$_{ds}$ curve shown in Fig. 4, describes that the variation of drain current with the drain to source voltage. The measured threshold voltage is approximately 0.3 V, and the maximum leakage current is approximate 3.5pA@accross all the corners (TT, SS, FF, SS125, and FF-40).

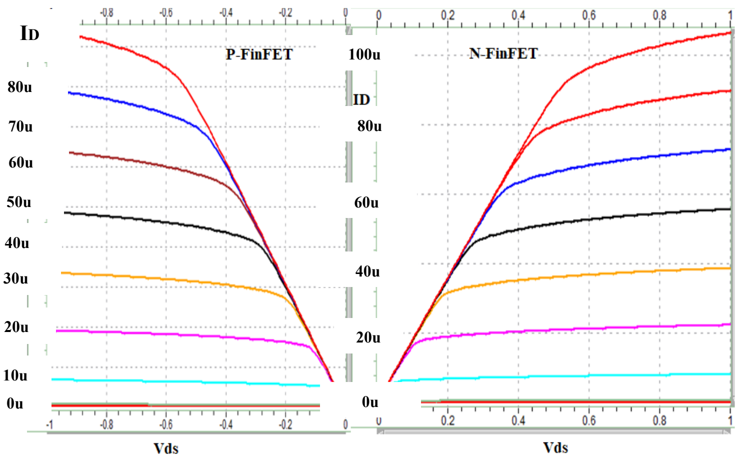


(a)

(b)

(c)

Fig. 3. I$_D$ vs V$_{gs}$ (a)Test Bench (b) P-FinFET (c) N-FinFET



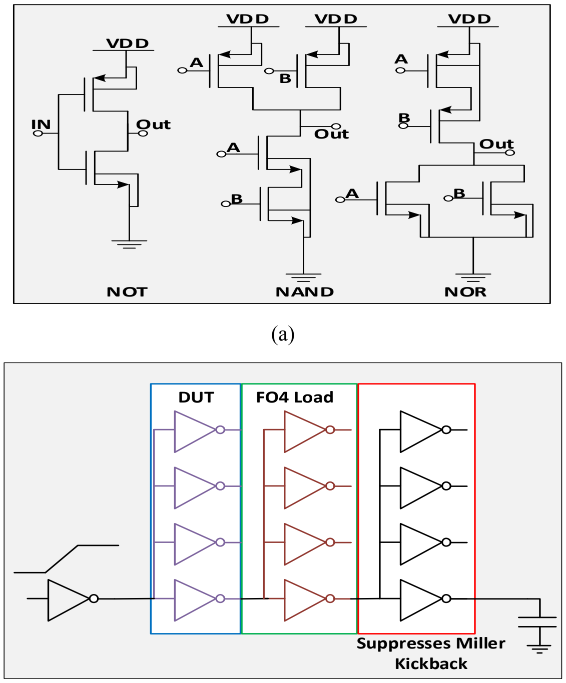
(a) (b)

Fig. 4. I_D vs V_{ds} (a) P-FinFET ID (b) N-FinFET

III. DIGITAL LOGIC IMPLEMENTATION USING THE FINFET

A. Estimation of the FO4 delay of basic Gates

In the present scenario, the junction and interconnect parasitic capacitance increase a lot, which impacted the drivability of the inverter as well as other logic gates. Therefore, FO4 delay or fan-out 4 delay is considered as the industrial standard. Here DUT stands for the device under test. The circuit diagram of the basic gates and FO4 test bench is shown in Fig 5(a) and (b), respectively [5] [6]. For the simulation 1V is used as supply voltage 45nm based circuit. Similarly, For the simulation 0.7V is used as supply voltage 7nm based circuit. The measured simulated FO4 delay of basic gate are compiled in the Table I. It is observed that the 7nm FinFET performed better than the 45nm CMOS technology.



(a)

(b)

Fig. 5. (a) Circuit diagram of the basic gates (b)FO4 Test Bench Setup

TABLE I. COMPARISON TABLE.

Gates	FO4 Delay 45nm	FO4 Delay 7nm
Inverter	10ps	6.873ps
NAND	13.55ps	8.06655ps
NOR	13.86ps	7.957ps
XOR	25.78ps	11.083ps
XNOR	21.02ps	10.329ps

B. Design of 16 Bit Adder

The block level representation of the 16-bit Brent Kung parallel prefix adder (PPA) is shown in Fig. 6 and the working principle are explained in reference [6]. Similarly, a ripple carries adder (RCA) is implemented by using a full adder as mention in reference [5] [6]. The carry propagation path is considered as the critical delay path therefore the FO4 delay is measured on the carry path. The simulation results are compiled in Table II.

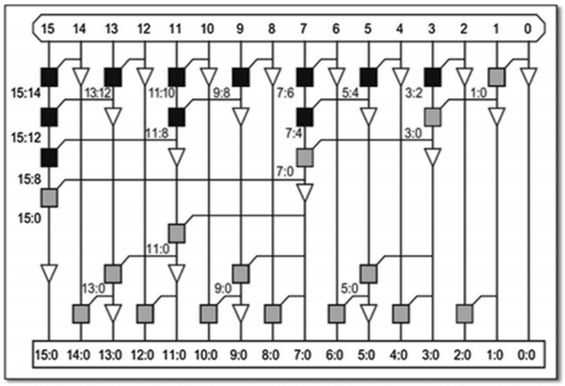


Fig. 6. 16-bit Brent Kung parallel [6]

TABLE II. COMPARISON TABLE.

4GHz 16-bit Brent Kung adder (PPA) and RCA				
Logic	Power	FO4 Delay	Energy	Energy*Delay
Static PPA 45nm	390.2uW	106.4ps	41.517fJ	4.417-24JS
Static RCA 45nm	343.3uW	333.ps	114.4fJ	38.14-24JS
Static PPA 7nm	38.05uW	58.82ps	2.238fJ	2.23e-24JS
Static RCA 7nm	31.14uW	167.2ps	5.206fJ	5.206e-24JS

C. 4GHz 16-Bit Pipeline array Multiplier

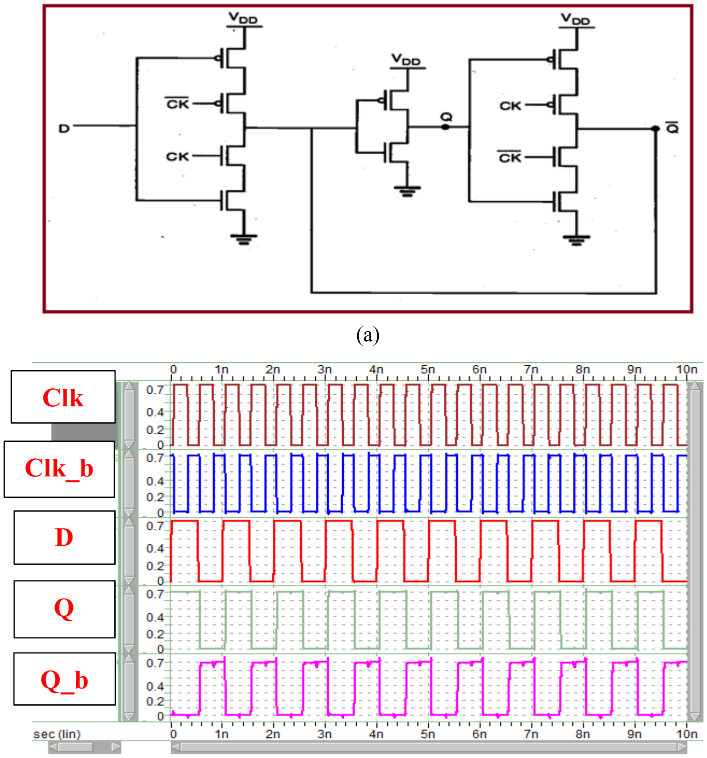
The circuit methodology proposed by Noll et al. (1986) is used herein to implement the 4GHz-16-bit pipelined arrayed multiplier [7]. The sub-block of the multiplier is designed using the full adder, D-FF and the basic logic gates. The circuit diagram and the simulated output of the D latch is shown in Fig 7 (a) and (b) respectively [5]. The performance of the design latch is shown in Table III. The performance comparison of the designed multiplier is shown in Table IV.

TABLE III. COMPARISON TABLE.

Delay	Power	Energy
0.5ns	0.2839uW	0.1446fJ

TABLE IV. COMPARISON TABLE.

4GHz Pipeline Array Multiplier				
Configuration	Supply	Power	Energy	No. of stage
Static Multiplier 45nm	1V	3.06mW	1.24pJ	16
Static Multiplier 7nm	0.7	0.9mW	0.36pJ	16



(a)

(b)

Fig. 7. Latch circuit used in the Multiplier and the output waveform

D. 6T SRAM CELL [8]

Fig. 8 shows the circuit diagram of the 6T static random-access memory (SRAM) cell [6][8]. For the CMOS technology, the main design constraint is that the size (W/L) of the driver transistor (M1 and M2) must be greater than the size of the write access transistor (A1 and A2). On the other hand, the size of the write access transistor must be greater than the pull transistor (M4 and M3) [9].

On the other hand, with the FinFET technology, the NFIN of the driver (M1 and M2) is three times as compared to the Pullup transistor (M4 and M3). Similarly, the NFIN of the driver (M1 and M2) is two times the NFIN of the access transistor (A1 and A2). Simulation results for the Hold Static Noise margin (HSNM), Read static Noise margin (RSNM) & Write static Noise margin (WSNM) is compiled in Table III.

The butterfly curve for HSNM, RSNM & WSNM is shown in Fig.9, Fig.10, and Fig. 11 respectively. Despite the reduced supply voltage in the 7nm node, the noise margin of the SRAM is still comparable to 45nm.

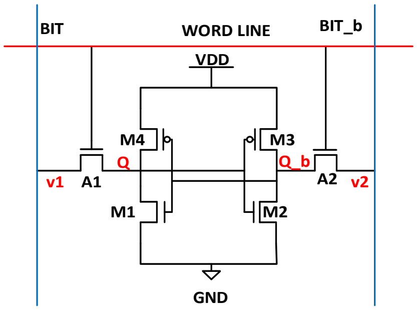


Fig. 8. Circuit diagram of the 6T SRAM

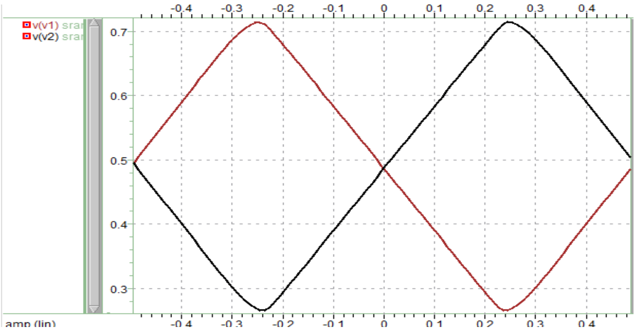


Fig. 9. Butterfly curve for HSNM

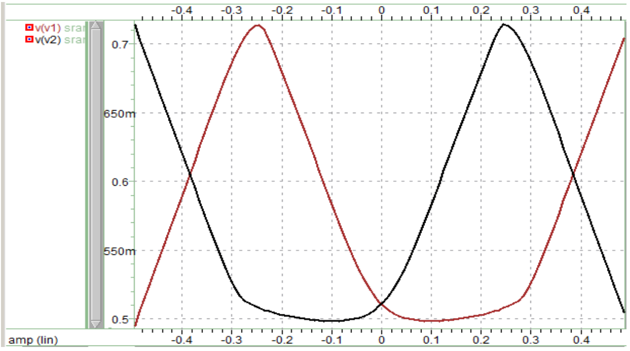


Fig. 10. Butterfly curve for Read SNM

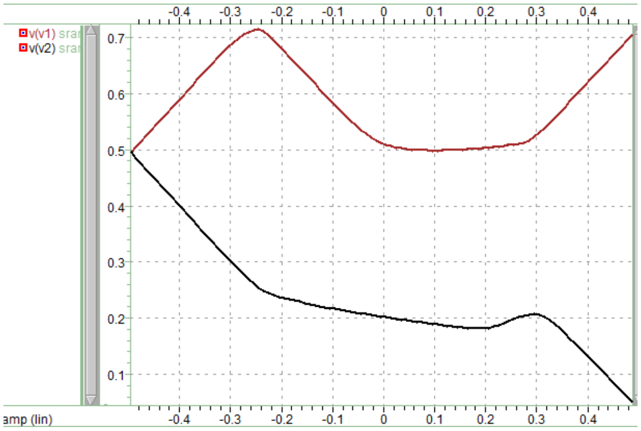


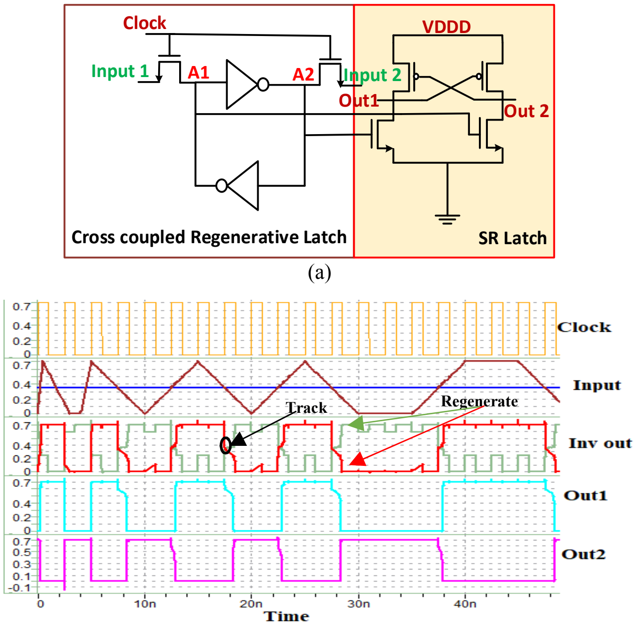
Fig. 11. Butterfly curve for Write SNM

TABLE V. COMPARISON TABLE.

6T SRAM			
Technology	Hold SNM	Read SNM	Write SNM
45nm	0.3306V	0.1787V	0.33V
7nm	0.3170V	0.1548V	0.4619V

E. Comparator Design using 7nm node

The circuit diagram of the regenerative latch comparator is shown in Fig. 12 (a). It incorporates a cross-coupled inverter and SR latch. Trans-conductor is designed herein by using the cross-coupled inverter. It is needed because the regeneration needs positive feedback, which can be easily implemented with the cross-coupled FinFET inverter.



(a)

(b)

Fig. 12. Regenerative latch with SR latch and output waveform

Regeneration node (A1& A2) is the output of the cross-coupled inverter. The input to the circuit charged the regeneration node with the help of N-FinFET. The input control switches are clocked at the frequency of the 4GHz. Therefore, the regenerative latch circuit works in two phases named as Track and Regeneration phase. When the clock is high, the input transistor is "ON" and regeneration node parasitic capacitance charges. When the clock signals low, the node which has the higher charge, regenerate toward the logic high, and another node goes towards the low. Since the output of the cross-coupled regenerative latch is mixed type, therefore SR latch stage is used herein to latch the output node to the full digital level. The sizing strategy is similar to the SRAM and the simulated output waveform is shown in Fig. 12(b). The total measured power consumption is 0.1mW @ 4GHz.

IV. CONCLUSION

The characterization of the N-FinFET and P-FinFET shows that the drain current of the N/P FinFET is directly proportional to the number of fin (NFIN) and the fingers. The measured threshold voltage is approximately 0.3V, and the measured maximum leakage current is 3.5pA across all the corners. During the simulation, it observes that the threshold voltage (0.3V) and the drain current of the N/P FinFET are almost equal. In order to avoid the timing and power specification violation during the synthesis of the NN, the prior knowledge of the delay and power is required. In this study the basic gates, DFF,16-bit Brent-Kung adder, 16-bit array pipeline multiplier and comparators are considered as the basic building block of discrete time NN implementation and subsequently the FO4 delay and power each block have been reported. Simulation results shows that the designed circuit can work efficiently with the clock frequency of 4GHz. The successful design of the 6T Static Random-Access memory (SRAM) and the clocked regenerative latch-based comparator (clocked at 4GHz) shows that the high-speed memory and mixed-signal circuit design is possible with 7nm FinFET. Despite the reduced supply voltage in the 7nm node, the noise margin of the SRAM is still comparable to 45nm. Performance comparison across various logic styles shows that the logic design using 7nm is highly energy efficient as compared to older (45nm) nodes.

ACKNOWLEDGMENT

Author would like thank TSRI Taiwan for the EDA tools supports. The authors want to thanks Professor Wei Hwang, MIRC, NCTU for teaching digital integrated circuit design course.

REFERENCES

[1] X. Yao, "Evolving artificial neural networks," Proceedings of the IEEE, vol. 87, no. 9, pp. 1423-1447, 1999.

[2] Khandelwal, S., Duarte, J.P., Medury, A., Venugopalan, S., Paydavosi, N., Lu, D.D., Lin, C., Dunga, M., Yao, S., Morshed, T.H., Niknejad, A., Salahuddin, S., & Hu, C. (2014). BSIM-CMG 110.0.0: Multi-gate MOSFET compact model: technical manual.

[3] Y. S. Chauhan, D. D. Lu, V. Sriramkumar, S. Khandelwal, J. P. Duarte, N. Payvadosi, A. Niknejad, and C. Hu, "FinFET Modeling for IC Simulation and Design: Using the BSIM-CMG Standard," Academic Press, 298 pages, 2015.

[4] HSPICE® Reference Manual: MOSFET Models Version J-2014.09, September 2014

[5] Sung-Mo (Steve) Kang and Yusuf Leblebici. 2002. CMOS Digital Integrated Circuits Analysis &Amp; Design (3 ed.). McGraw-Hill, Inc., New York, NY, USA.

[6] Neil H. E. Weste and Kamran Eshraghian. 1985. Principles of CMOS VLSI Design: a Systems Perspective. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[7] T. G. Noll, D. Schmitt-Landsiedel, H. Klar and G. Enders, "A Pipelined 330-MHz Multiplier," in IEEE Journal of Solid-State Circuits, vol. 21, no. 3, pp. 411-416, June 1986, doi: 10.1109/JSSC.1986.1052543.

[8] W. S. Hsu et al., "28nm ultra-low power near-/sub-threshold first-in-first-out (FIFO) memory for multi-bio-signal sensing platforms," 2016 International Symposium on VLSI Design, Automation and Test (VLSI-DAT), Hsinchu, 2016, pp. 1-4.

[9] E. Seevinck, F. J. List and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," in IEEE Journal of Solid-State Circuits, vol. 22, no. 5, pp. 748-754, Oct 1987.

[10] R. K. Pandey and S. k. Pandey, "High Resolution Comparator Design for RF Imager," 2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS), Thiruvananthapuram, India, 2018, pp. 65-69.