INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2017, VOL. 63, NO. 4, PP. 369–374
Manuscript received January 10, 2017; revised September, 2017. DOI: 10.1515/eletel-2017-0050

Ultra Low Power Design for Digital CMOS Circuits Operating Near Threshold

Shruti Kalra, Amalendu B. Bhattacharyya

***Abstract*—Circuits operating in the subthreshold region are synonymous to low energy operation. However, the penalty in performance is colossal. In this paper, we investigate how designing in moderate inversion region recuperates some of that lost performance, while remaining very near to the minimum energy point. An α power based minimum energy delay modeling that is continuous over the weak, moderate, and strong inversion regions is presented. The value of α is obtained through interpolation following EKV model. The effect of supply voltage and device sizing on the minimum energy and performance is determined. The proposed model is utilized to design a temperature to time generator at 32nm technology node as the application of the proposed model. The abstract goes here.**

***Keywords*—Energy Efficiency, Ultra Low Power, EKV, Minimum Energy point, Minimum Delay point, temperature to time generator**

I. Introduction

VLSI scaling has proceeded throughout the last few decades, empowering reasonable and productive devices which led advanced lives. There were a few difficulties on the way undermining progress: design efficiency in the 80's, power utilization in the 90's, and leakage in the last decade [1]. In spite of the fact that the technology scaling kept continued, multiplying transistors in each era or reducing the supply voltage does not decrease the energy per operation to use every one of the transistors. In this way, the next challenge is of energy efficiency:- not simply low power but to keep on delivering rationale throughput with a less energy utilization. Energy considerations are essential over almost all the applications range including high-execution platforms to low power circuits. While Moore's law allows more number of transistors, just a part may be utilized at once because of power restrictions and performance of the application in this manner is gagged by power limits, frequently in the 500 mW to 5W range[5][6]. The designer in this period is to defeat the test of energy efficient computing and unleash performance from the reins of energy to re enable Moore's law in the semiconductor business. Subthreshold operation is perfect for this class of circuits since it permits minimum energy operation for low-performance. But, the penalty for working at the minimum energy point is significant. The delay at the minimum energy point is around three orders bigger than at minimum delay point. To reduce energy consumption, voltage scaling technique has demonstrated a promising technique with subthreshold outline representing endpoint of voltage scaling. Despite the fact that it is to a great degree energy proficient, subthreshold design has been consigned to corner markets because of its real performance punishments[8]. A large amount articles are available in the literature that focuses on determining the minimum energy point but reliability of any circuit at lower supply voltage is a matter for research, since many digital circuits fail to switch when the supply is reduced or process variations and becomes more susceptible to noise[5][6][10].

Our proposed procedure is to give ten times or higher energy efficiency at consistent performance and widespread applications of near threshold computing where devices are operated at or close to their threshold voltage. By reducing supply voltage from an ostensible 1V to 300 to 400 mV, near threshold circuits acquires as much as ten times energy efficiency and represents the restoration of voltage scaling and its associated energy gains at ultradeep submicron technology node. This paper characterizes and investigates near threshold computing, a design space where the supply voltage is around equivalent to the threshold voltage of the transistors. This inversion holds a lot of the energy savings of subthreshold operation with more positive performance and variability characteristics[10]. The highlights of the model are:

- Low power analytical design of digital CMOS circuits is largely based on approach followed by [2] where a simplified long channel current equation is used with velocity saturation index $\alpha = 2$ and a fitting parameter kfit to take care of the velocity saturation effect. This assumption however is too restrictive because the above mentioned approach is applicable for the condition where inversion coefficient IC=1 when gate to source voltage V_{gs} is equal to the threshold voltage V_{th}. However, the range of IC varies from 1 to 10 and for each operating condition kfit has to be adjusted. Therefore, a simplified α power law based model for MOS transistor as a function of voltages that fits the moderate and weak inversion regions is derived for ultradeep submicron technology node. The model matches especially well the behavior around the threshold till 32nm technology node with industry standard BSIM4 and is less complex as compared to the model of [4].
- In our approach, the value of α is uniquely defined for each IC and no other fitting parameter is required for delay as against [2]. The value of α is derived from interpolation as done by EKV model [3].

Authors are with the Department of Electronics and Communication, Jaypee Institute Of Information Technology, Noida, India (e-mail: shruti.kalra@jiit.ac.in, ab.bhattacharyya@jiit.ac.in).

- Further, delay and energy models that can be utilized as a part of logic analysis tool are determined and verified using BSIM4 simulation.
- Energy-delay sensitivity to sizing, supply and threshold and energy-delay trade-off for inverter driving another inverter at 32nm technology node is determined
- The proposed model is utilized to design a temperature to time generator at 32nm technology node as the application of the proposed model.

The paper is organized as follows: Section II describes the near threshold drain current model. Section III highlights the gate delay model for near threshold computing. Energy model is presented in section IV. Sensitivity analysis of supply voltage, device sizing and threshold voltage on energy delay trade-off is given in section V. Section VI applies the proposed model to time-to-digital based smart temperature sensor circuit. In the end, the entire work is concluded in the conclusion section.

II. Modeling the Near Threshold Current

The basic conventional model isolates MOSFET operation into three unique modes (subthreshold, linear and saturation), depending upon the voltages at the transistor terminals. This presents a trouble in analyzing circuits operating near threshold since the model does not contain a solitary, consistent equation for the near threshold operation, i.e. where VGS<VT and VGS>VT. The EKV model [3] was appeared to be extremely accurate for current in the near threshold locale for submicron technologies. The model uses a mathematical interpolation with a specific end goal to associate the separate regions of operation into a one model, accordingly giving an exact expectation when the MOSFET is working in the near threshold region. Based on the EKV formulas [3],the drain current I_{ds} can be expressed as:

$$I_{ds} = I_S.IC \tag{1}$$

Here,I_S is the specific current equal to $I_S = 2n\mu_{eff}C_{ox}\phi_t^2 W/L$. n is the subthreshold slope, C_{ox} is the oxide capacitance, μ_{eff} is the effective mobility, ϕ_t is the thermal voltage, W is the channel width and L is the effective channel length. IC represents the inversion coefficient. The inversion coefficient expresses the degree of inversion of the transistor, and covers both the sub-$V_{th}(IC < 1)$ and above $V_{th}(IC > 1)$ regions. Now,

$$i_{ds} = ln^2\left[1 + exp\left(\frac{v_p - v_s}{2}\right)\right] - ln^2\left[1 + exp\left(\frac{v_p - v_d}{2}\right)\right] \tag{2}$$

where $IC = I_{ds}/I_S = i_{ds}$ is the normalized value of drain current to make it unitless, $v_p = \frac{V_{gs}-V_{th}}{n\phi_t}$ is the normalized value of pinch-off voltage and $v_d = V_d/\phi_t$ and $v_s = V_s/\phi_t$ is the normalized value of drain and source voltage respectively and V_{th} is the threshold voltage. Referencing the above equation through the source instaed of bulk and removing the normalization,

$$I_{ds} = I_s ln^2\left[1 + exp\left(\frac{V_{gs} - V_{th}}{2n\phi_t}\right)\right] - I_s ln^2\left[1 + exp\left(\frac{V_{gs} - V_{th}}{2n\phi_t} - \frac{V_{ds}}{2\phi_t}\right)\right] \tag{3}$$

ON current of MOSFET, I_{ON} is defined as drain current when $V_{gs} = V_{ds} = V_{dd}$. Thus, equation 3 becomes:

$$I_{ON} = I_s ln^2\left[1 + exp\left(\frac{V_{dd} - V_{th}}{2n\phi_t}\right)\right] - I_s ln^2\left[1 + exp\left(\frac{V_{dd} - V_{th}}{2n\phi_t} - \frac{V_{dd}}{2\phi_t}\right)\right] \tag{4}$$

Assuming the supply voltage V_{dd} to be few times larger than ϕ_t, the second term in the above expression becomes negligible. Thus,

$$I_{ON} = I_s ln^2\left[1 + exp\left(\frac{V_{dd} - V_{th}}{2n\phi_t}\right)\right] \tag{5}$$

The model in 5 is based on the EKV formulas and is for long channel device [3]. For a short channel device operating at low supply voltage, [2] proposes a model by introducing drain current and delay fitting parameters, kfit and ktp respectively. But the model is restricted to inversion coefficient IC=1. But, for moderate inversion region where the supply voltage is low, the value of IC ranges from 1 to 10. . Therefore, Modification to the above model is proposed by introducing a parameter α as:

$$I_{ON} = I_s ln^{\alpha}\left[1 + exp\left(\frac{V_{dd} - V_{th}}{2n\phi_t}\right)\right] \tag{6}$$

The value of α can be obtained using interpolation of weak and strong inversion.

The leakage current at gate to source voltage $V_{gs} = 0$ with reverse saturation current I_S can be expressed as [3]:

$$I_{Leakage} = I_S exp\left(\frac{V_{dd} - V_{th}}{n\phi_t}\right) \tag{7}$$

III. Modeling the Near Threshold Delay

By utilizing the drain current model described in the above section, an expression for delay is determined for a CMOS inverter. To begin with, consider the instance of discharging the output capacitance with NMOS, the propagation delay t_p of a gate can be expressed as:

$$t_p = \frac{C_L V_{dd}}{I_{ON}} \tag{8}$$

Here, C_L is the load capacitance. I_{ON} is the ON current of NMOS. Putting the value of I_{ON} from 6

$$t_p = \frac{C_L V_{dd}}{2n\mu_{eff}C_{ox}\frac{W}{L}{\phi_t}^2}\frac{1}{IC} \tag{9}$$

Here,

$$IC = ln^{\alpha}\left[1 + exp\left(\frac{V_{dd} - V_{th}}{2n\phi_t}\right)\right] \tag{10}$$

Load capacitance C_L corresponds to the sum of intrinsic capacitance from the driving stage and load capacitance of the fan-out gates. Thus C_L is proportional to:

$$C_L \propto C_{ox}L(\xi_i W_i + W_{i+1}) \tag{11}$$

Here, ξ is the ratio of parasitic capacitance of the driver stage and input gate capacitance of the fanout. i and i+1 are the annotations for driver and load stage respectively. Thus, W in equation 9 becomes W_i. Putting the value C_L in equation 9,

$$t_p = \frac{L^2}{2n\mu_{eff}{\phi_t}^2}\frac{V_{dd}}{IC}\frac{\xi_i W_i + W_{i+1}}{W_i} \tag{12}$$

For N number of stages along the path, the path delay can be approximated as:

$$t_{p,i} = \frac{L^2}{2n\mu eff{\phi_t}^2}\frac{V_{dd}}{IC}\sum_{i=1}^{N}\frac{\xi_i W_i + W_{i+1}}{W_i} \tag{13}$$

$$t_{p,i} = k_{pd}\frac{V_{dd}}{IC}t_{p,i}(W) \tag{14}$$

Here, k_{pd} is the technology dependent factor and $t_{p,i}(W)$ is depends on transistor size. In nutshell, delay for N stages can be often expressed in terms of average propagation delay t_p of single stage and the logic depth Ld.

$$t_{p,i} = t_p Ld$$

Figure 1 shows the variation of delay with supply voltage ranging from weak to moderate to strong inversion. The results obtained from analytical calculations are verified with BSIM4 for 32nm technology node. It can be seen from the plot that minimum delay is obtained at larger supply voltage.

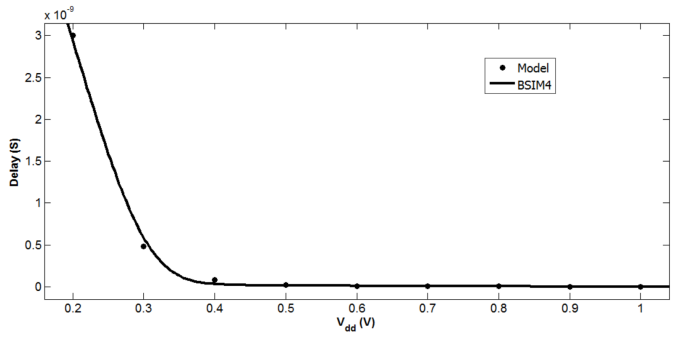

Fig. 1. Variation of propagation delay for minimum size CMOS inverter driving another inverter as a function of supply voltage. The analytical results are verified using BSIM4 at 32nm technology node.

IV. Modeling the Near Threshold Energy

The total energy e_{total} consumed by any logic gate is the sum of the switching energy i.e. energy corresponding to the charging and discharging of the load capacitance C_L, $e_{switching}$, and the leakage component of energy consumption, $e_{leakage}$, in one clock period. For a single gate, this can be expressed as follows:

$$e_{total} = e_{switching} + e_{leakage} \tag{15}$$

$$= C_L{V_{dd}}^2 a + TI_S exp\left(\frac{V_{dd} - V_{th}}{n\phi_t}\right)V_{dd} \tag{16}$$

Where T is the clock period which is assumed to be equal to $t_{p,i}$ and a is the activity factor. Since C_L is the function of channel width, putting the value of C_L from the equation 11:

$$e_{total,i} = C_{ox}L(\xi_i W_i + W_{i+1}){V_{dd}}^2 a + (k_{pd}\frac{V_{dd}}{IC}t_p(W))(2n\mu_{eff}C_{ox}\frac{W_i}{L}{\phi_t}^2)exp\left(\frac{V_{dd} - V_{th}}{n\phi_t}\right)V_{dd} \tag{17}$$

The energy consumed over N stages or path energy can be expressed by summation as:

$$e_{total,i} = C_{ox}L\sum_{i=1}^{N}(\xi_i W_i + W_{i+1}){V_{dd}}^2 a + (k_{pd}\frac{V_{dd}}{IC}t_p(W))(2n\mu_{eff}C_{ox}\frac{\sum_{i=1}^{N}W_i}{L}{\phi_t}^2)exp\left(\frac{V_{dd} - V_{th}}{n\phi_t}\right)V_{dd} \tag{18}$$

Figure 2 shows the variation of total energy consumed over

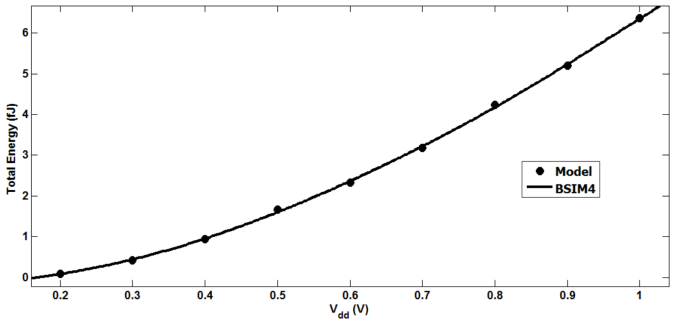

Fig. 2. Variation of energy consumed over one cycle for minimum size CMOS inverter driving another inverter as a function of supply voltage. The analytical results are verified using BSIM4 at 32nm technology node.

one period with supply voltage ranging from weak to moderate to strong inversion. The results obtained from analytical calculations are verified with BSIM4 for 32nm technology node. It can be seen from the plot that minimum energy is consumed at smaller supply voltage.

Energy consumption in present day CMOS circuits to a great extent results from the charging and discharging of load capacitance and can be reduced quadratically by bringing down the supply voltage V_{dd} as can be seen from the figure 2. All things considered, voltage scaling has become one of the more compelling techniques to lower power consumption. It is surely understood that the CMOS circuits remains functional at low supply voltage below the threshold voltage known as subthreshold voltages but the lack of robustness and poor performance limits the designer to operate the circuit in this region.

In the above threshold regime, energy is exceedingly sensitive to supply voltage V_{dd} because of the quadratic scaling. Thus, voltage scaling upto near threshold region yields energy reduction of around ten times at the cost of around ten times performance degradation as shown in figure 3. Nonetheless, the reliance of energy on supply voltage V_{dd} turns out to be more complex as voltage is scaled below the threshold voltage V_{th}. In subthreshold region, circuit delay increments exponentially with supply voltage V_{dd}, bringing about leakage energy i.e the leakage current, delay and supply voltage to

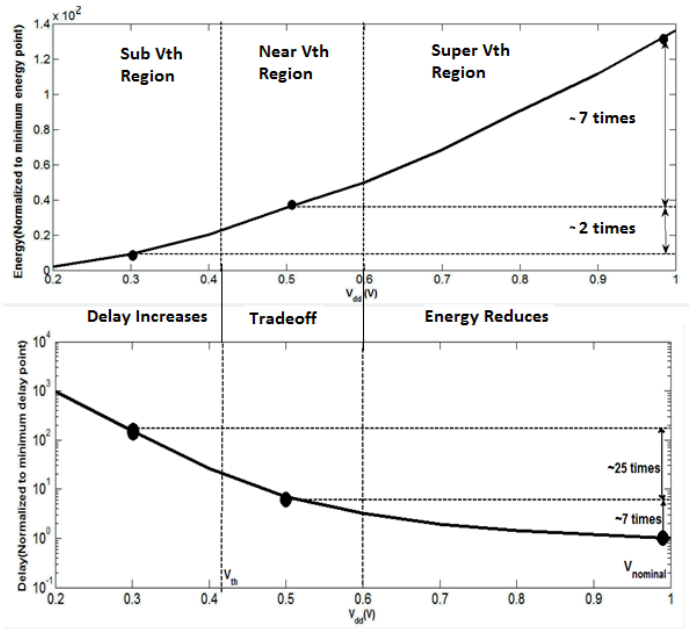

Fig. 3. Energy and delay in different supply voltage operating regions for analysing energy/delay trade-off for minimum size CMOS inverter driving another inverter at 32nm technology node.

increment in a close exponential manner. This increase in leakage energy will stop reduction in any energy and creates minimum energy point.

V. Sensitivity Analysis

In this section, the effect of device sizing, supply and threshold voltage has been analyzed on tradeoff between delay and energy. The tradeoff by means of voltage and device sizing will be evaluated utilizing the idea of energy-delay sensitivity. The sensitivity to any parameter p shows the percent decrease in energy for a percent increment in delay represented as:

$$S(P) = \frac{\frac{\delta(e_{total,i})}{\delta P}}{\frac{\delta(t_{p,i})}{\delta P}} \tag{19}$$

To show the sensitivity of supply voltage V_{dd} and channel width W_i in the energy-delay space, figure 4 shows the plot of energy delay space when the supply voltage V_{dd} and channel width W_i are exclusively tuned, beginning from minimum energy point. As anticipated, scaling the supply voltage is more effective than scaling the channel width W_i around minimum energy point. The channel width scaling is barely compelling until we are near to minimum delay point. Along these lines, dissimilar to minimum delay point where channel width scaling was the most predominant optimization variable, supply voltage ought to be utilized around minimum energy point. This is on the grounds that at minimum energy point, leakage current varies linearly with device width W_i and so does the circuit performance. The supply voltage V_{dd} scaling is more viable near minimum energy point because of the fact that supply voltage exponentially influences performance.

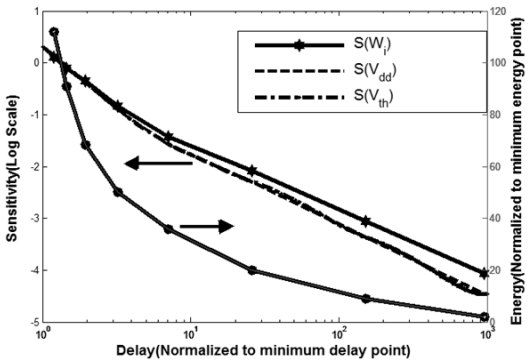

Fig. 4. Energy-delay sensitivity to device sizing, supply voltage and threshold voltage for minimum size CMOS inverter driving another inverter at 32nm technology node.

Given the substantial divergence in device sizing and supply sensitivities, we may reduce device sizing (if conceivable) around minimum energy point that creates energy slack and can be used by a small increment in supply voltage for general performance increment. This is comparable, though in distinctive order of changing variables, to increase the supply voltage around minimum delay point that creates timing slack and can be used by device sizing for reducing energy. These types of trade-offs are by and large unrealistic at minimum energy point or minimum delay point since the device sizing and supply variables achieve their limits, so the utilization of device sizing near minimum delay point or supply voltage near minimum energy point is ideal.

VI. Temperature to Time Generator

The block diagram of temperature to time generator and its timing diagram is shown in figure 5 [11]. Delay line 1 comprises of multiple temperature compensated delay cells of low thermal sensitivity. In contrast, delay line 2 is made from delay buffers that are not temperature compensated. Toward the start of every estimation, the start signal is postponed by delay lines 1 and 2 individually to get the postponed signals A and B. A succeeding XOR is utilized to create the delay difference between two delay and will be fed to the input of the time to digital converter [11].

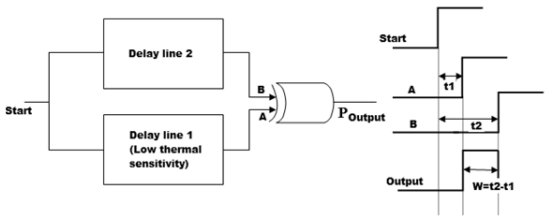

Fig. 5. Circuit of the temperature-to-time generator.

The propagation delay for the CMOS inverter driving another inverter can be expressed by equation 12. The temperature dependence of various parameters of MOSFET are identified and described in [12]. The temperature dependent

parameters for MOSFET in equation 12 are: I_S,V_{th},ϕ_t and α.

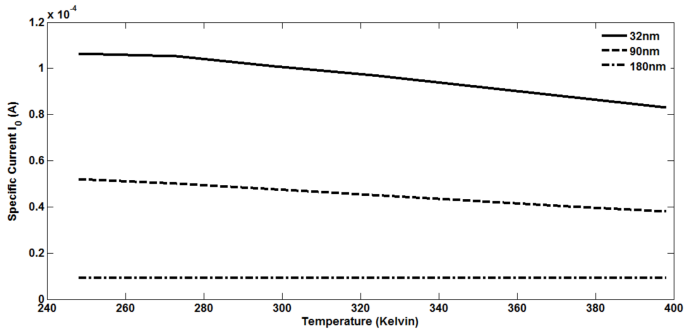

Fig. 6. Variation of the specific current I_0 with temperature for different technology nodes. The plot have been curve fitted to the quadratic relationship given in equation 20

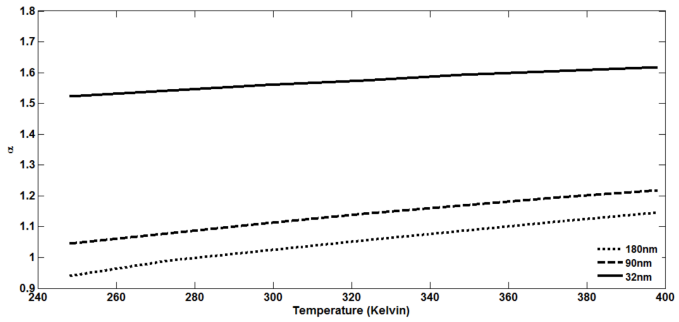

Fig. 7. Velocity Saturation Index α dependence on temperature (25-125 Degrees) for three technology nodes for the transistor operating in moderate inversion

These parameters vary with temperature as [12]:

$$I_S(T) = I_0(T) = A(T)^2 + B(T) + C \tag{20}$$

$$\alpha(T) = D(T)^2 + E(T) + F \tag{21}$$

$$V_{th}\,(T) = V_{th}\,(T_0) - \alpha_T \frac{T}{T_0} \tag{22}$$

Here A,B,C,D,E,F are the fitting parameters, $V_{th}\,(T_0)$ is the threshold voltage at room temperature and α_T is the temperature coefficient of threshold voltage. The value of α and I_0 was extracted at three technology nodes on various temperatures and curve fitted to quadratic function as shown in the figure 6 and 7 and given in equation 20 and 22. By putting the various parameter variations in the delay equation, we could easily find the variation of propagation delay with temperature. The higher the temperature is, the higher is the propagation delay of delay line 2. The delay line 1 with lower thermal sensitivity is utilized to reduce the offset at the lower bound of the measurement range which could otherwise cause longer conversion time. The XOR gate is being used to produce delay difference between the two delay lines. The width of the output pulse from the XOR gate can be controlled by number of stages of delay line 2. Figure 8 shows the temperature compensated delay cell utilized for delay line 1. This thermal compensation circuit is used to reduce the sensitivity of the inverters used in delay line 1. The inverters used are the same as that used in delay line 2. Here the diode-connected transistors P1, N1, and P3 serve as the core of the thermal compensation circuit.

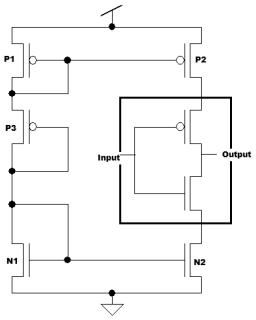

Fig. 8. Temperature compensated delay cell utilized for delay line 1

Since P1, P3, and N1 are all diode connected, they will operate in saturation if bias current is flowing. Thus, from equation 6:

$$I_{ds,P3}(T) = I_s(T) ln^{\alpha(T)} \left[1 + exp\left(\frac{V_{dd} - V_{th}(T)}{2n\phi_t(T)}\right)\right] \tag{23}$$

For a diode connected MOSFET, in order to get minimum thermal sensitivity,

$$\frac{dI_{ds,P3}}{dT} = 0$$

Putting the temperature dependencies from equations 20, 21 and 22 and differentiating the equation 23 to find the temperature insensitive value of V_{dd} for different technology nodes. The temperature insensitive value of supply voltage $V_{dd} = 0.35V$ for 32nm technology node and lies in the moderate inversion region. Figure 9 shows the analytically calculated pulse width of temperature insensitive delay line 1, temperature sensitive delay line 2, the output (P_{output}) of XOR gate. The results obtained are verified using industry standard BSIM4. It can be seen from the figure that the offset at the lower bound of the measurement range is reduced as against when only delay line 2 is used. This reduces the conversion time.

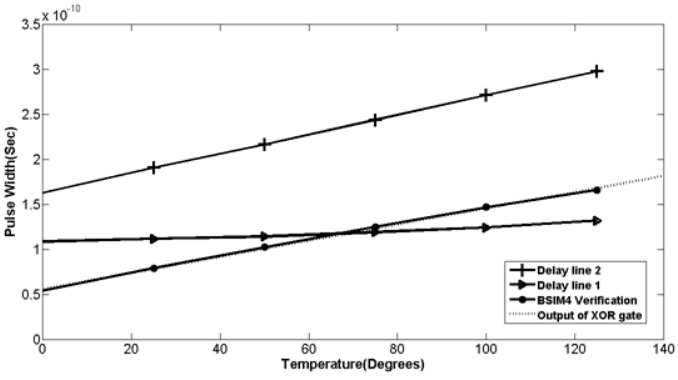

Fig. 9. Pulse width of temperature insensitive delay line 1, temperature sensitive delay line 2 and the output (P_{output}) of XOR gate

VII. Conclusion

As Moore's law keeps on providing circuit designers with additional transistors on a chip, power costs are starting to constrain the pertinence of these extra transistors in CMOS circuit design. In this paper we glanced back at the attainability of voltage scaling to reduce energy utilization. In spite of the fact that subthreshold operation is surely understood to give generous energy savings, it has been consigned to a small number of applications because of the relating degradation of the performance. Thus, the idea of near threshold processing has been explored, where the supply voltage is at or close to the threshold voltage of the transistors. It has been demonstrated that using interpolation, for drain current in the moderate inversion, the velocity saturation index α can be determined for ultradeep submicron technology node. This administration empowers energy savings around ten times, with just around ten times performance degradation, giving a much better tradeoff in energy/delay than subthreshold operation. With conventional device scaling giving no more energy proficiency improvements, our essential decision is that the answer for this energy problem is the general use of near threshold operation. As against the various near threshold computing models available in the literature, the proposed model is applicable in the ultradeep submicron technology node. The results obtained at 32nm technology node has been analyzed and verified through BSIM4 simulations and the error lies within the acceptable range of 5 to 10 $\%$.

VIII. Non referenced citations

[7], [9]

References

[1] Dreslinski, Ronald G., Michael Wieckowski, David Blaauw, Dennis Sylvester, and Trevor Mudge. "Near-threshold computing: Reclaiming moore's law through energy efficient integrated circuits." Proceedings of the IEEE 98, no. 2 (2010): 253-266.

[2] Markovic, Dejan, Cheng C. Wang, Louis P. Alarcon, Tsung-Te Liu, and Jan M. Rabaey. "Ultralow-power design in near-threshold region." Proceedings of the IEEE 98, no. 2 (2010): 237-252.

[3] Enz, Christian C., and Eric A. Vittoz. Charge-based MOS transistor modeling: the EKV model for low-power and RF IC design. John Wiley & Sons, 2006.

[4] Keller, Sean, David Money Harris, and Alain J. Martin. "A compact transregional model for digital CMOS circuits operating near threshold." Very Large Scale Integration (VLSI) Systems, IEEE Transactions on 22, no. 10 (2014): 2041-2053.

[5] Kaul, Himanshu, Mark Anders, Steven Hsu, Amit Agarwal, Ram Krishnamurthy, and Shekhar Borkar. "Near-threshold voltage (NTV) design: opportunities and challenges." In Proceedings of the 49th Annual Design Automation Conference, pp. 1153-1158. ACM, 2012.

[6] Chen, Gregory, Dennis Sylvester, David Blaauw, and Trevor Mudge. "Yield-driven near-threshold SRAM design." IEEE transactions on very large scale integration (VLSI) systems 18, no. 11 (2010): 1590-1598.

[7] Calhoun, Benton H., Alice Wang, and Anantha Chandrakasan. "Modeling and sizing for minimum energy operation in subthreshold circuits." IEEE Journal of Solid-State Circuits 40, no. 9 (2005): 1778-1786.

[8] Alioto, Massimo. "Ultra-low power VLSI circuit design demystified and explained: A tutorial." IEEE Transactions on Circuits and Systems I: Regular Papers 59, no. 1 (2012): 3-29.

[9] Gonzalez, Ricardo, Benjamin M. Gordon, and Mark A. Horowitz. "Supply and threshold voltage scaling for low power CMOS." IEEE Journal of Solid-State Circuits 32, no. 8 (1997): 1210-1216.

[10] Taur, Yuan. "CMOS design near the limit of scaling." IBM Journal of Research and Development 46, no. 2.3 (2002): 213-222.

[11] Chen, Poki, et al. "A time-to-digital-converter-based CMOS smart temperature sensor." IEEE Journal of Solid-State Circuits 40.8 (2005): 1642-1648.

[12] Kalra, Shruti, et al. "An Analytical Study Of Temperature Dependence of Scaled CMOS Digital Circuits Using a-Power MOSFET Model" Journal of Integrated Circuits and Systems 11.1 (2016): 57-68.