Design of a Quasi-Linear Rail-to-Rail Delay Element with an Extended Programmable Range

Jordan Lee Gauci*, Edward Gatt*, Owen Casha*, Giacinto De Cataldo†, Ivan Grech* and Joseph Micallef*
*Department of Microelectronics and Nanoelectronics, University of Malta
†Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Italy
*E-mail: jordan-lee.gauci.10@um.edu.mt

***Abstract*—This paper presents an analytical model of a quasi-linear delay element to be used in the High Momentum Particle Identification Detector (HMPID) at the CERN Large Hadron Collider (LHC). The aim of this model is to facilitate the design of a delay element in order to achieve the required range while maximizing linearity. In addition, a technique is proposed to further increase the delay range by means of a programmable banked capacitor architecture without sacrificing linearity. The design of a rail-to-rail quasi-linear delay element with a range spanning from $125\,\mathrm{ns}$ to $580\,\mathrm{ns}$ was achieved. The proposed model was verified via simulations performed in Cadence using the X-FAB $0.18\,\mu\mathrm{m}$ technology.**

***Index Terms*—Delay lines, Delay element, wide-range, linearity, modelling**

I. Introduction

Precise delay generation is an important research area as delay generators can be found in a number of applications ranging from high-energy physics to time-based analog-to-digital converters. The delay line is at the heart of the delay generator and the delay can be controlled either via analog or digital means. Various architectures have been proposed such as the current-starved inverter architecture and the diode-connected transistor architecture [1]. These architectures have their merits and disadvantages, however it is generally difficult to obtain a wide delay range while achieving a linear response with a rail-to-rail operation. For instance, the work in [2] proposed a 1.8 V highly linear delay element, featuring a delay range of $0.5\,\mathrm{ns}$ to $4.5\,\mathrm{ns}$. In this case, the analogue tuning voltage was limited to $0.9\,\mathrm{V}$.

This paper presents the design of a rail-to-rail quasi-linear wide range delay element. The design is aided by means of an analytical model based on the delay element circuit proposed in [3]. The presented model facilitates the sizing of the transistor aspect ratios in order to obtain the required range and linearity without requiring lengthy parametric simulations. In addition, this work proposes a technique to further increase the range, while maintaining a symmetric operation. This work will be used in the High Momentum Particle Identification Detector (HMPID) at the CERN Large Hadron Collider (LHC). HMPID is a triggered detector, where the data stored on the charge pads is read upon the reception of a trigger signal, which drives the sample-and-hold circuitry. It is therefore essential that the charge on the pads is read at its maximal value, in order to obtain an optimal signal-to-noise ratio [4]. Currently, the trigger signal arrives approximately $1.2\,\mu\mathrm{s}$ after a collision has occurred, however after the second long shutdown period (2019-2021), HMPID will be utilizing another trigger signal that arrives after approximately $700\,\mathrm{ns}$. This will therefore require the use of a highly accurate delay generator such that the timing of the trigger signal can be fine-tuned. In addition, it is important to have a wide delay range with a linear and monotonic transfer characteristic.

II. Delay Element Architecture and Delay Model

The conventional current-starved inverter architecture suffers from a non-linear relationship between the delay and the tuning voltage while having a limited input range from V_t to V_{DD}, where V_t is the transistor threshold voltage and V_{DD} is the supply voltage. The delay time (T_d) between the input and the output signal of a current-starved inverter can be modelled by [3, 5]:

$$T_d \propto \frac{C_L}{I_{cp}} V_{DD} \tag{1}$$

where C_L represents the capacitive load of the inverter and I_{cp} is the charging and discharging current through the capacitive load. Eq. 1 shows that the delay may either be varied through the control of C_L, I_{cp}, or V_{DD}. Typically, only the current is varied, leading to a non-linear response.

The constant of proportionality of Eq. 1 depends on the value of the load capacitance and the charging/discharging current. For large capacitances, the discharge current would be limited by the sizing of the control transistors. This is equivalent to having a constant current discharging a capacitor. Assuming that since the discharge current is constant, at least initially, the voltage across the capacitor, V_{out}, decreases linearly. Thus, for the ideal case

$$I_{cp} = -C_L \frac{dV}{dt} \tag{2}$$

$$\int_0^{T_d} dt = \frac{C}{I_{cp}} \int_{\frac{V_{DD}}{2}}^{V_{DD}} dV \tag{3}$$

$$T_d = \frac{C_L}{I_{cp}} \frac{V_{DD}}{2} \tag{4}$$

In reality, however, the value of V_{out} does not start from V_{DD}, but from $V_{DD} + \delta V$, where δV is the contribution due to charge injection due to the parasitic capacitances between the gates and drains of the transistors forming the inverter.

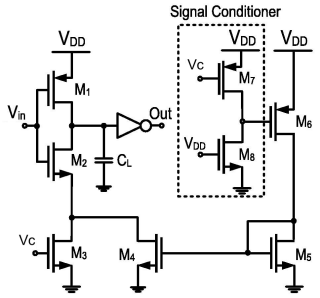

Fig. 1. Linear Delay Element Circuit [3].

This therefore implies that the expression of the time delay consists of two parts; the time it takes for the output voltage to reach V_{DD} from $V_{DD}+\delta V$, and the time to discharge from V_{DD} to $\frac{V_{DD}}{2}$. This would therefore imply that

$$T_d = \frac{C_L}{I_{cp}}\left(\frac{V_{DD}}{2} + \delta V\right) \tag{5}$$

The value δV may be obtained from

$$\delta V = \frac{C_p}{C_L} dV_{in} \tag{6}$$

where C_p is the total parasitic capacitance between gate and drain of the two inverter transistors, and dV_{in} is the change in input voltage until it reaches its final value.

The delay element based on the work proposed in [3], is illustrated in Fig. 1. The circuit is based on a current-starved inverter architecture and can obtain both a quasi-linear delay and rail-to-rail operation. This is achieved via transistors $M_4 - M_8$. If the tuning voltage, V_c, is applied directly via M_6, the delay response of the circuit would be highly non-linear and non-monotonic. Thus, an inverting common-source amplifier (consisting of M_7 and M_8) is used in order to achieve a monotonic and quasi-linear relationship in the delay response of the circuit.

The work in [6] showed that for a control voltage range $V_{tn} < V_c < V_{DD} - V_{tp}$, a linear delay transfer characteristic may be achieved by setting the aspect ratios of transistors M_3, M_5, and M_8 to 0.9, 20.8, and 3.9, respectively. Nonetheless, since the analytical model, used for the design and optimization, is not completely valid for rail-to-rail operation [6], simulations show that the linearity worsens for $V_c < V_t$. Furthermore, this circuit is able to delay only the falling edge of the input signal, thus limiting the delay generated by the pulse width of the input.

To overcome this issue, the circuit illustrated in Fig. 2 is being proposed. Transistors $M_2 - M_7$ form the delay element [3], which is controlled by the control voltage V_c. The current from this architecture is then mirrored to the balanced current-starved inverter, formed by transistors $M_{10} - M_{13}$. This results in a delay element that is capable of rail-to-rail operation with a quasi-linear response with an extended programmable range. The delay can be finely tuned via V_c and a coarse control is provided via switches EN_1 and EN_2,

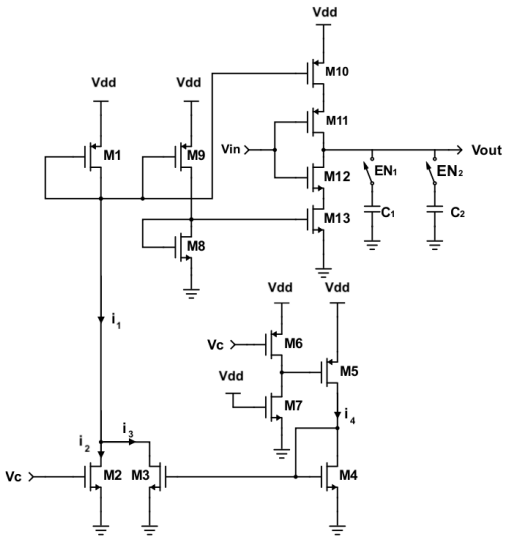

Fig. 2. Improved linear delay element circuit with extended programmable range and symmetric operation.

which increases the effective capacitive load at the output node. In addition, this circuit also enables an increase in the delay range via proper scaling of the current mirror ratios M_{10}/M_1 and M_{13}/M_8. This is particularly useful for limiting the size of the on-chip capacitors. For a symmetrical current-starved inverter, the charging and discharging current should be equal. This means that $i_{charge} = \frac{R10}{R1} i_1$ must be equal to $i_{discharge} = \frac{R13}{R8} i_1$. The aspect ratios of M_{11} and M_{12} should be as large as possible, in order for the charging and discharging currents to be fully controlled by M_{10} and M_{13}.

III. Analytical Model

An analytical model for the current in the circuit of Fig. 2, and subsequently the delay generated, is presented in this section, where the total load capacitance of the delay element is given by C_L. Assuming that M_2 and M_3 remain in pinch-off, the piece-wise expression for the current in M_1 is given by:

$$i_1 = \begin{cases} i_3 & V_c < V_{tn} \\ i_2 + i_3 & V_c \geq V_{tn} \end{cases} \tag{7}$$

where i_x is the current through transistor M_x and x is the transistor identifier. When M_2 and M_3 operate in pinch-off, the currents are given by:

$$i_2 = \frac{K'_n}{2}\frac{W_2}{L_2}(V_c - V_{tn})^2 \tag{8}$$

$$i_3 = \frac{\frac{W_3}{L_3}}{\frac{W_4}{L_4}} i_4 \tag{9}$$

$i_4 = i_5$ is given by

$$i_4 = \frac{K'_p}{2}\frac{W_5}{L_5}(V_{sg5} - V_{tp})^2 \tag{10}$$

where

$$V_{sg5} = V_N - \sqrt{(V_N)^2 - 2\frac{K_p}{K_n}(V_{DD} - V_c - V_{tp})^2} \tag{11}$$

and $V_N = V_{DD} - V_{tn}$, $K_p = K'_p \frac{W_6}{L_6}$ and $K_n = K'_n \frac{W_7}{L_7}$. This means that Eq. 7 may be represented as:

$$i_1 = \begin{cases} \alpha_3 V_c^2 + \alpha_2 V_c + \alpha_1 V_c^{1/2} + \alpha_0 & V_c < V_{tn} \\ \alpha_7 V_c^2 + \alpha_6 V_c + \alpha_5 V_c^{1/2} + \alpha_4 & V_c \geq V_{tn} \end{cases} \quad (12)$$

where parameters α_0 to α_7 are functions of the transistor aspect ratios, process parameters K'_n and K'_p, the supply voltage, and the threshold voltages of the NMOS and PMOS transistors. The delay that can be generated may still be modelled as a piece-wise equation of Eq. 5. The relationships involved are highly complex, making it difficult to design the delay cell by means of a closed form equation.

To be able to simplify this equation, an approximation technique based on the Newton Polynomial method was employed. This technique allows its transformation into a piece-wise second-order polynomial with good accuracy. Therefore, the time delay equation can now be approximated by:

$$T_d \approx \begin{cases} A_1 V_c^2 + B_1 V_c + C_1 & V_c < V_{tn} \\ A_2 V_c^2 + B_2 V_c + C_2 & V_c \geq V_{tn} \end{cases} \quad (13)$$

where A_x, B_x, C_x are model parameters depending on the transistor aspect ratios, process parameters, supply voltage, and the threshold voltages. A closer examination of the coefficients in Eq. 13 show that the linearity depends mainly on the aspect ratios of transistors M_6 and M_7 while the range depends on the aspect ratio of transistors M_3, M_4 and M_5. Once the desired range has been obtained, the aspect ratio of M_2 was chosen, such that its effect would not limit the linearity to the range $V_{tn} \leq V_c \leq V_{DD}$. This method results in a quasi-linear relationship between the delay and control voltage while ensuring the desired range. With this polynomial, the aspect ratios of the transistors may be optimised to obtain a quasi-linear response without requiring lengthy parametric simulations.

IV. Verification of Model and Simulations

The circuit shown in Fig. 2 was implemented and simulated in Cadence using the $0.18\,\mu\text{m}$ X-FAB technology. The optimized transistor aspect ratios are presented in Table I. To limit channel length modulation effects, the minimum length size was chosen as $500\,\text{nm}$. A $1\,\text{pF}$ double metal-insulator-metal capacitor (cdmm4) was used and covers an area of $20 \times 20\ \mu\text{m}^2$. This capacitor was chosen because it provides the largest capacitance per unit area ($2\,\text{fF}/\mu\text{m}^2$) and it has a small voltage coefficient of 3 ppmV^{-1}.

For the analytical model to take into consideration any effects related to transistor mismatches, each transistor was characterized in order to obtain the different parameters K'_n and K'_p together with the respective threshold voltages. The delay was calculated by applying a pulse to the V_{in} input, and calculating the time it takes for the output to reach $0.5V_{dd}$, for different control voltages V_c. The Newton Polynomial method was used to approximate the delay equation by two piece-wise second-order polynomials. The results are illustrated in Fig. 3, where the analytical model is plotted in black, the approximation model is plotted in blue and the results obtained from the simulator in red. It can be seen that through this method, a good approximation may be obtained, through which the transistor aspect ratios may be found in order to obtain the most linear delay. Since a piece-wise second-order polynomial describes the response of the delay element, the coefficient of V_c^2 can be minimised for the case when $V_c < V_{tn}$, where there is no dependency on the aspect ratio of M_2. M_2 can then be chosen such that its effect for the case when $V_{tn} \leq V_c \leq V_{DD}$, is minimized as much as possible. This will ensure that the delay response remains quasi-linear.

TABLE I
Transistor aspect ratios of the linear delay generator.

Transistor Name	Aspect Ratio
M_2	0.1
M_3	12
M_4	12
M_5	1
M_6	5.9
M_7	4.85

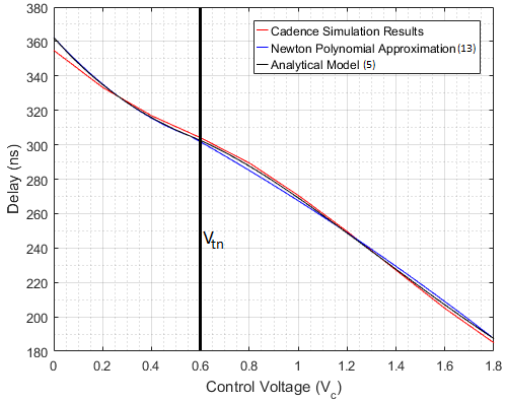

Fig. 3. Comparison of the simulated results with those obtained from the analytical model and the Newton Polynomial approximation method.

To further test the linearity of the circuit, a sinusoidal input, with a frequency of $2.148\,\text{kHz}$ was applied to V_c, and an input square wave of $200\,\text{kHz}$ was applied to V_{in}, with a sampling frequency of $2\,\text{GHz}$, and a simulation time of $5.12\,\text{ms}$. The delay between the input and the output was then calculated via a MATLAB script. The spectrum of the delay is plotted in Fig. 4. The Spurious-Free Dynamic Range (SFDR) of the delay is equal to $25.11\,\text{dB}$, and the Signal-to-Noise-and-Distortion-Ratio is equal to $23.04\,\text{dB}$. Although these values are less than those achieved in [3], the proposed delay element circuit has a wider delay range ($168\,\text{ns}$ in this work compared to $1.4\,\text{ns}$ in [3]).

To confirm that the delay element with extended range works as expected, a $500\,\text{fF}$ capacitor, and a $1\,\text{pF}$ capacitor, were used. Each of the capacitors was connected to a transmission gate to be able to include or exclude them from the circuit. This allows for an increase in the delay range. The results are presented in Fig. 5 for the different combinations of the capacitors. As expected, since the capacitance is multiplied by

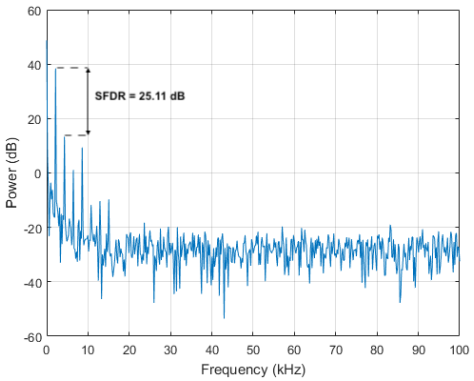

Fig. 4. Frequency spectrum of the simulated time delay signal.

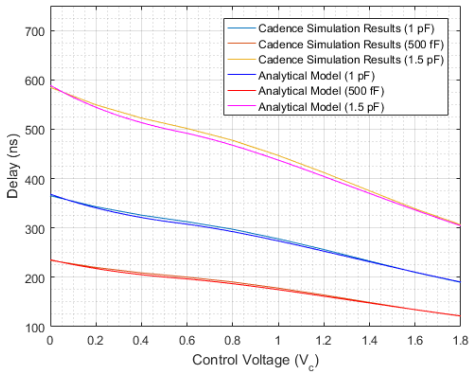

Fig. 5. Extended delay range achieved via the proposed programmable banked capacitor architecture.

the value of the inverse of the current, then both the gradient and the y-intercept will change. This is because the value of the capacitance is multiplied by the inverse of the current, which approximates a straight line.

Temperature analysis and Monte Carlo analysis were carried out to test the robustness of the delay element for $C_L = 1\,\mathrm{pF}$. The results from the temperature analysis are presented in Fig. 6, where it can be observed that the delay varies with the operating temperature. Linearity suffers at temperatures in the range of $-40\,^{\circ}\mathrm{C}$ to $-20\,^{\circ}\mathrm{C}$, particularly near the threshold voltage of M_2. Monte Carlo simulations were performed for process and mismatch variations with 200 points for all values of V_c. The results show that for $V_c = 0\,\mathrm{V}$ the mean delay between the rising edge at the input and the falling edge at the output has a mean delay of $374.3\,\mathrm{ns}$ with a standard deviation of $51.61\,\mathrm{ns}$. On the other hand, the mean delay between the falling edge at the input and the rising edge at the output has a mean delay of $370.3\,\mathrm{ns}$ with a standard deviation of $38.8\,\mathrm{ns}$. This shows that there is a slight discrepancy between the charge and discharge paths of the current due to process variations and that the process variations affect more the NMOS side than the PMOS side. This is further evident when the results from mismatch contributions are taken into consideration. In fact, the largest error contribution comes from transistors M_3, and M_{13}, at 39% each. M_8 and process variations are the next largest contributors at 35% each.

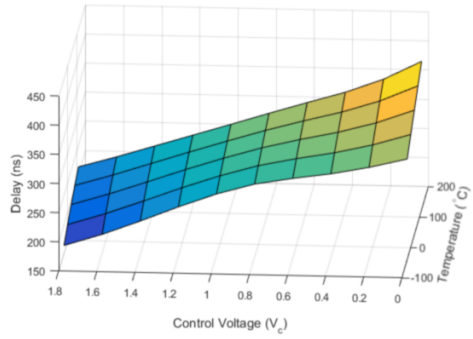

Fig. 6. Simulated delay variation with control voltage across a temperature sweep from $-40\,^{\circ}\mathrm{C}$ to $120\,^{\circ}\mathrm{C}$.

V. Conclusion

This paper has presented the design of a rail-to-rail quasi-linear wide range delay element as an improvement to the current-starved inverter proposed in [3]. In addition, this work has also proposed a method to obtain a wider delay-range, through the use of a programmable banked-capacitor architecture while maintaining symmetric operation. An analytical model was proposed as an aid for the design. The developed model was approximated through the use of the Newton Polynomial method, which facilitates the sizing of the transistor aspect ratios, in order to achieve a quasi-linear delay transfer characteristic. The model was validated across process and temperature variations via simulation using the X-FAB $0.18\,\mu\mathrm{m}$ technology.

Acknowledgements

The research work disclosed in this publication is funded by the ENDEAVOUR Scholarship Scheme (Malta). The scholarship may be part-financed by the European Union - European Social Fund (ESF) under Operational Programme II - Cohesion Policy 2014-2020, *"Investing in human capital to create more opportunities and promote the well being of society."*

References

[1] B. Abdulrazzaq, I. A. Halin, S. Kawahito, R. M. Sidek, S. Shafie, and N. A. Yunus, "A Review on High-Resolution CMOS Delay Lines: Towards sub-picosecond Jitter Performance," *SpringerPlus*, vol. 5, no. 1, pp. 1–32, 2016.

[2] A. Seraj, M. Maymandi-Nejad, and M. Sachdev, "A new linear delay element with self calibration," in *23rd IEEE Iranian Conference on Electrical Engineering (ICEE)*, 2015, pp. 1050–1053.

[3] H. Rivandi, S. Ebrahimi, and M. Saberi, "A low-power rail-to-rail input-range linear delay element circuit," *AEU-International Journal of Electronics and Communications*, vol. 79, pp. 26–32, 2017.

[4] J. L. Gauci, E. Gatt, G. De Cataldo, O. Casha, and I. Grech, "An Analytical Model of the Delay Generator for the Triggering of Particle Detectors at CERN LHC," in *2017 IEEE New Generation of CAS (NGCAS)*, 2017, pp. 69–72.

[5] E. Zafarkhah, M. Maymandi-Nejad, and M. Zare, "Improved accuracy equation for propagation delay of a CMOS inverter in a single ended ring oscillator," *AEU-International Journal of Electronics and Communications*, vol. 71, pp. 110–117, 2017.

[6] J. L. Gauci, E. Gatt, O. Casha, G. De Cataldo, and I. Grech, "On the Design of a Linear Delay Element for the Triggering Module at CERN LHC," in *14th IEEE Conference on PhD Research in Microelectronics and Electronics (PRIME)*, 2018.