CMOS Technology Scaling Advantages in Time Domain Signal Processing

Jussi-Pekka Jansson, Pekka Keränen, Juha Kostamovaara, *Senior Member, IEEE*
and Andrea Baschirotto, *Fellow, IEEE*

***Abstract*—This paper compares two CMOS technologies, the robust 350nm version and its modern 28nm successor, in terms of time-domain signal processing parameters. The evaluated parameters; propagation delay, delay variation due to process and mismatch fluctuations, sensitivity to noise and area and power usage are crucial especially in measurement devices relying on precise timings, high precision time-to-digital converters, for example. Post-layout simulations show that the modern scaled technology offers superior speed, efficient area usage and low power consumption but suffers from considerable delay mismatch. Therefore applications relying on precise time domain signal processing do not always benefit from technology scaling.**

***Index Terms*—Technology scaling, time-domain, CMOS, delay element, ring oscillator, time-to-digital converter, TDC.**

I. Introduction

SCALED CMOS technology has been developed to enable progressively more efficient electronic blocks and systems. Better manufacturing accuracy and smaller development dimensions have made it possible to integrate ever larger numbers of transistors and structures into the same circuit, yielding more and more sophisticated single-chip systems and smaller-sized electronic devices. These scaled technologies also mean a reduction in the operating voltage, and their internal capacitance lowers power consumption, which benefits mobile electronics and reduces cooling requirements.

In terms of signal processing performance, technology scaling towards smaller development dimensions has both pros and cons for the different IPs to be developed. For example, smaller transistor dimensions and denser structures shorten the signal propagation time, allowing for high-performance digital signal processing, e.g. high frequency oscillators, while technology scaling in general benefits fast, low-power, large-scale digital signal processing. On the other hand, such scaling weakens some of the transistor properties which are important for analogue signal processing (such as output impedance, (V_{DD}-V_{TH}) distance and parasitic components). In fact, signal dynamic range and intrinsic gain decrease with technology scaling, and compensatory structures (often referred as digital-assisted techniques) need to be considered when the "analogue" properties of the technology are not sufficient [1], [2].

Precise phase matching or phase differences in the time domain will often be a required feature in high-speed integrated circuits. Oscillators, clock trees and frequency synthesis, for example, can benefit from constant or predictable delays between signals. One of the most demanding applications from this point of view is the Time-to-Digital-Converter (TDC), which aims at assigning a precise digital value for the interval between two or more timing signals. TDCs are important blocks in many measurement devices and applications in nuclear science, location, laser scanning and perception etc.

This paper analyses the effects of technology scaling on time domain signal processing parameters. Modern (28nm) and older (350nm) CMOS technologies are compared in terms of propagation delay and the delay variation caused by process parameter fluctuations, mismatch and noise. These are the most important parameters as far as signal integrity in the time domain is concerned and have a crucial influence on the measurement performance of TDCs, for example [3].

The theoretical model for inverter delay and its variation reveals first the factors affecting the delay and its behaviour. Simulations carried out based on parasitic extracted layout structures will serve then to describe the difference between the two technologies. These simulations evaluate a delay-adjustable buffer-type cell suitable for use as a ring oscillator or a high precision TDC. The effects of fluctuating technology parameters on TDC measurement performance are estimated before reaching conclusions.

II. Signal propagation delay and its variation

A. Propagation delay

The exact propagation delay is difficult to analyse even for the simplest CMOS logic gate, an inverter, due to the markedly non-linear input-to-output characteristics. Its device-related capacitances are non-linear, and short-channel effects make the analysis even more difficult in modern CMOS technologies. Although the exact value for the propagation delay cannot be accurately derived, the following analysis aims to clarify how the device dimensions affect it and the mismatch.

In general, the delay depends on the capacitive load and the drive-strength of the inverter, in connection with which Fig. 1 illustrates the capacitances associated with an inverter having another inverter as a load. In general, the capacitances consist of a coupling capacitance, C_C, from input-to-output, and a load capacitance, C_L, to GND/V_{DD}.

Manuscript received October 31. The work has been supported financially by the Academy of Finland and INFN (ScalTech28 experiment), which are gratefully acknowledged.

J.-P. Jansson, P. Keränen and J. Kostamovaara are with the Circuits and Systems Research Group, University of Oulu, Oulu 90014, Finland. (e-mail: jussi.jansson@oulu.fi, pekka.keranen@oulu.fi, juha.kostamovaara@oulu.fi).

A. Baschirotto is with the Department of Physics, University of Milano-Bicocca, Milano 20126, Italy. (e-mail: andrea.baschirotto@unimib.it).

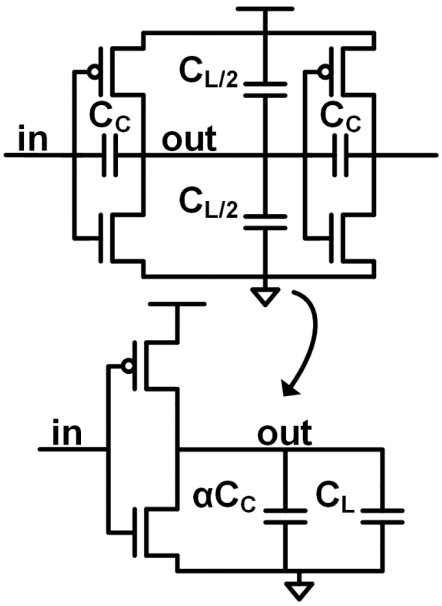

Fig 1. Capacitances affecting the delay in an inverter.

Simulations show that increasing the coupling capacitance increases the delay in a linear manner, so that the coupling capacitance can be replaced by a load capacitor with a value of αC_C, where α is a fitting parameter which, accounting for the Miller effect, for example.

The coupling capacitance incorporates an overlap capacitance proportional to the transistor width, *W*, of the nmos/pmos. For simplicity, the transistors of the driving inverter are assumed to be either in cut-off or in saturation, so that the gate-channel capacitance to the drain is zero. The coupling capacitance is then

$$C_C \approx C_{OLp}W_p + C_{OLn}W_n \ , \tag{1}$$

where C_{OLp} and C_{OLn} are the overlap capacitances of the pmos/nmos.

The load capacitance consists of the wiring capacitance, the gate-channel (to-source) capacitances of the subsequent inverter, the overlap capacitances of the subsequent inverter and the drain-to-bulk junction capacitance of the driving inverter.

$$C_L \approx C_{wire} + C_{Jp}W_p + C_{Jn}W_n + C_{OLp}W_p + C_{OLn}W_n + C_{ox}(W_n + W_p)L \ , \tag{2}$$

where C_{Jp}/C_{Jn} is the drain-to-bulk junction capacitance and C_{ox} is the gate-channel capacitance to source. The total load capacitance is

$$C_{tot} = \alpha C_C + C_L \ . \tag{3}$$

In order to simplify the equation for the total load capacitance, it is assumed that the overlap capacitances and the junction capacitances for both devices are roughly the same. Furthermor, to achieve equal delays for high-to-low and low-to-high transitions, it is assumed that the width, W_p, of the pmos is *k* times larger than the width, W_n, of the nmos. The length, *L*, is equal for both devices. The effective load capacitance can be now written as

$$C_{tot} \approx (k+1)W_n(C_{ox}L + (1+\alpha)C_{OL} + C_J) + C_{wire} \ . \tag{4}$$

The exact delay in an inverter is rather difficult to analyse due to non-linearities. We assume here for simplicity that the input is a voltage step from *GND* to V_{DD} and the pmos is in cut-off and the nmos in saturation. When a voltage step is applied at the input at a time t=0, the output voltage of the inverter is given by:

$$V_{out}(t) = V_{dd} - \frac{I_{DSAT}}{C_{tot}}t \ . \tag{5}$$

The propagation delay. i.e. the time for the output voltage to reach $V_{DD}/2$, is given roughly by:

$$\tau \approx \frac{C_{tot}}{I_{DSAT}}\frac{V_{dd}}{2} = \frac{\left((k+1)W_n(C_{ox}L + (1+\alpha)C_{OL} + C_J) + C_{wire}\right)V_{dd}}{2K_n\frac{W_n}{L}\left((V_{dd} - V_{Tn})V_{DSAT} - \frac{V_{DSAT}^2}{2}\right)} \ . \tag{6}$$

where K_n is the current factor. The equation suggests that it is possible to reduce the inverter propagation delay by increasing W_n, but in actual fact a larger value for W_n reduces the effect of C_{wire} and the propagation delay saturates to a certain level τ_{min} independent of W_n.

$$\tau_{min} \approx \frac{(k+1)L(C_{ox}L + (1+\alpha)C_{OL} + C_J)V_{dd}}{2K_n\left((V_{dd} - V_{Tn})V_{DSAT} - \frac{V_{DSAT}^2}{2}\right)} \ . \tag{7}$$

The transistor length *L* has a strong impact on the propagation delay and should be minimized for high speed signal processing. Furthermore, when velocity saturation is taken into account, i.e. $V_{DSAT} = Lv_{sat}/\mu_n$, the minimum propagation delay becomes

$$\tau_{minvsat} \approx \frac{(k+1)\mu_n(C_{ox}L + (1+\alpha)C_{OL} + C_J)V_{dd}}{2K_nv_{sat}\left((V_{dd} - V_{Tn}) - \frac{Lv_{sat}}{2\mu_n}\right)} \ . \tag{8}$$

B. Propagation delay variation

A fabricated circuit naturally has variations in the process parameters and in the processed layout. Transistor width and length, threshold voltage, oxide thickness and wiring, for example, may introduces mismatches between devices designed to be identical and lead to differences in w.r.t. nominal values. Because of this, the logic cell propagation delay also varies with PVT cases.

The effects of local random variations with respect to the current factor and threshold can be deduced by differentiating the absolute delay in the inverter with respect to K_n and V_{Tn}.

The load capacitance variations have not been taken into account in this analysis, because in the general Pelgrom model [4] variation parameters are only available for the current factor and the threshold voltage. Small changes in K_n and V_{Tn} result in small variations in the propagation delay as follows:

$$\begin{aligned}\Delta\tau_{K_n} &= \Delta K_n\left(\frac{d}{dK_n}\tau\right)\\ &= -\frac{\Delta K_n}{K_n}\frac{C_{tot}V_{dd}}{2K_n\frac{W_n}{L}\left((V_{dd}-V_{Tn})V_{DSAT}-\frac{V_{DSAT}^2}{2}\right)}\\ &= -\frac{\Delta K_n}{K_n}\tau\ .\end{aligned} \quad (9)$$

$$\begin{aligned}\Delta\tau_{\Delta V_{Tn}} &= \Delta V_{Tn}\left(\frac{d}{dV_{Tn}}\tau\right)\\ &= \frac{\Delta V_{Tn}V_{DSAT}C_{tot}V_{dd}}{2K_n\frac{W_n}{L}\left((V_{dd}-V_{Tn})V_{DSAT}-\frac{V_{DSAT}^2}{2}\right)^2}\\ &= \frac{\Delta V_{Tn}}{V_{dd}-V_{Tn}-\frac{V_{DSAT}}{2}}\tau\ .\end{aligned} \quad (10)$$

Consequently the standard deviation for the delay variation is given by

$$\sigma_{\Delta\tau} = \frac{\tau}{\sqrt{W_nL}}\sqrt{\sigma_{\frac{\Delta K_n}{K_n}}^2+\sigma_{\Delta V_{Tn}}^2\frac{1}{\left(V_{dd}-V_{Tn}-\frac{V_{DSAT}}{2}\right)^2}}\ , \quad (11)$$

where $\sigma_{\frac{\Delta K_n}{K_n}}^2$ and $\sigma_{\Delta V_{Tn}}^2$ are technology-dependent local variance parameters for the current factor and the threshold voltage. As can be seen in (11), the delay variation depends on the propagation delay, device sizes and technology-dependent parameters.

III. Technology comparison with post-layout simulations

The delay-adjustable digital cell shown in Fig. 2 will be used as the benchmark structure in the technology comparison analysis. Differential opposite input signals arrive at the cell input (IN+ and IN-), altering the values of the corresponding opposite output signals (OUT1- and OUT1+). Small cross-coupled inverters between the outputs, in the middle, correct any phase error between the opposite output signals. The secondary output inverters (OUT2+) and (OUT2-) act as an extra load for the cell. The propagation delay is adjusted with the control voltage V_{CTRL}, which limits the current when the signals propagate through the cell. The number next to each transistor describes its W/L ratio, which is altered in the simulations by means of a multiplier M.

The experiments were performed by embedding the cell in a ring oscillator structure and, for better accuracy, the comparison analysis was performed on post-layout simulations. The cell layout was drawn for each simulation case presented in Table I and parasitic capacitances and resistances were extracted. The layouts were also similar for both technologies, only the transistor widths and lengths being changed.

The measured parameters were expressed in absolute (ps or fs) and relative (%) values, the absolute values indicating how much the measured parameters were in one delay element in real time, while in the relative expressions the absolute values were divided by the corresponding propagation delay τ_p. This reveals in general how much the measured parameter affects the time domain signal processing.

A. Propagation delay

The propagation delay τ_p through the cell, as presented in Fig. 2, was measured between the signals IN+ and OUT+. V_{CTRL} was adjusted in all the simulations so that the cell delay was 1.2 times the minimum delay. The process, voltage and temperature fluctuations also force one to use a propagation delay higher than the minimum in real applications. the cell delay at this biasing point is described by τ_p.

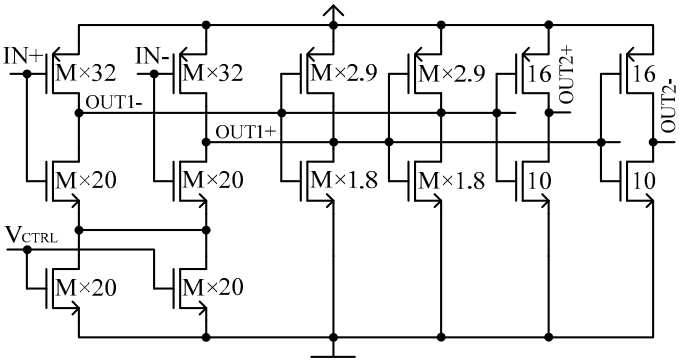

Fig. 2. Structure of a typical delay-adjustable delay element.

TABLE I
POST-LAYOUT SIMULATION RESULTS.

Simulation case	1	2	3	4	5	6	7	8	9
CMOS technology [nm]	350	350	350	350	28	28	28	28	28
Gate length L [nm]	350	350	350	350	28	28	28	28	56
Supply voltage [V]	3.3	3.3	3.3	3.3	1	1	1	1	1
W/L multiplier M	1	2	4	8	2	4	8	16	8
Propagation delay τ_p [ps]	101.9	86.0	77.2	74.4	14.4	12.5	11.6	11.6	17.8
$\sigma_{PROCESS}$ [ps]	12.3	10.4	9.2	9.1	1.2	1.0	0.95	1.0	1.1
$\sigma_{PROCESS}$ [%]	12.1	12.1	11.9	12.2	8.3	8.2	8.2	8.5	6.2
$\sigma_{MISMATCH}$ [fs]	1050	660	420	290	750	540	320	210	330
$\sigma_{MISMATCH}$ [%]	1.0	0.8	0.5	0.4	5.2	4.3	2.8	1.8	1.9
Δ_{DELAY} [ps] /10mV	0.6	0.5	0.5	0.4	0.3	0.2	0.2	0.2	0.4
Δ_{DELAY} [%] /10mV	0.6	0.6	0.6	0.5	2.1	1.6	1.7	1.7	2.2
Power consumption [mW]	13.2	23.8	44.9	86.5	0.8	1.5	2.8	5.2	3.6
Area [μm^2]	690	1000	1630	2890	30	45	70	130	150

The post-layout simulations show that the propagation delay τ_p decreases for larger values of M (the W/L ratio multiplier), but that this decrease slows down with larger transistor sizes, because the role of the permanent load, C_{wire}, diminishes, as predicted in (6). Case 9 shows that transistor L (gate length) has a strong effect on the delay, so that it should be minimized when a small propagation delay is desired. An increase in the length of a transistor will make its transconductance smaller and next-stage input capacitance higher, which will slow down the propagation delay, as predicted in (7). Comparing between the technologies shows that the 28nm CMOS technology offers superior speed, the propagation delay improves by about 6.2x when the technology is scaled by 12.5 (same W/L).

B. *Propagation delay variation*

Monte Carlo simulations (N=200) were performed to evaluate the effect of process variations ($\sigma_{PROCESS}$) and mismatch variations ($\sigma_{MISMATCH}$) on the propagation delay. $\sigma_{PROCESS}$ in Table 1 defines the expected standard deviation of the delay element propagation delay when the elements are created in different process runs. Using the minimum transistor length, $\sigma_{PROCESS}$ is about 12% (350nm CMOS) and 8.5% (28nm CMOS) of the mean value. Increasing W has no effect on this, but case 9 shows that increasing L improves the relative process mismatch. In other words, L should be increased when longer delays are needed with small delay variation.

$\sigma_{MISMATCH}$ describes the propagation delay variation (standard deviation) when the cells are located close to each other on the same wafer. The absolute value of $\sigma_{MISMATCH}$ decreases faster as the transistor width is increased, as predicted in (11). The first reason for this is that the absolute delay τ_p becomes smaller, and the second reason is that the effects of differences and imperfections are averaged in larger transistors (higher W in this case) and have less influence on the total delay. A defect in the transistor gate, for example, will have more effect on the transistor properties if the gate is small. Increasing L (case 9) increases the absolute mismatch because the delay becomes greater, but the relative mismatch remains the same. Hence increasing L is justified when larger delays need to be created without affecting the mismatch.

Comparison of the two technologies shows that the absolute value of $\sigma_{MISMATCH}$ is about the same when the same W/L value (M) is used. The relative mismatch is hence much higher in the scaled technology, where the propagation delay is smaller. The technology-dependent parameters (11) seem to be higher in the scaled technology, where the transistor size has been reduced in order to achieve high speed, a small area and low power consumption. It is also possible that the defects and imperfections may have become smaller but not to the same extent as the size. The smaller dimensions together with the weaker intrinsic gain bring out the small differences in transistor structure and create more variation in the delay.

C. *Sensitivity to noise*

The variation in propagation delay because of noise in the control voltages was estimated by changing the control voltage V_{CTRL} 10mV from the original biasing point. The blocks designed for a certain propagation delay are often delay-controlled with a control voltage, which is the most sensitive place for error. The effect of a similar voltage change in the supply is much smaller.

Variations in M (W/L ratio) will have no significant effect on the delay change Δ_{DELAY}. The relative delay change reveals that a similar variation in control voltages will have more effect on propagation delays in scaled than in older technologies. The intended propagation delay deviated from the original value by ~0.6% with 350nm CMOS and ~1.8% with 28nm CMOS, when the control voltage was changed by 10mV. It is mainly the smaller operating voltage in the scaled technology that makes the effects of noise and voltage errors more relevant.

D. *Power consumption and area*

The power consumption was estimated when the delay elements constituted an oscillating ring oscillator. The power consumption increased dramatically as M is increased. The delay element input capacitances increased, as also did the oscillating frequency, which had a considerable effect on the power consumption, which was about 30x higher with the 350nm technology than with the 28nm technology when oscillators with the same value of M were compared.

The area of the delay element naturally also almost doubles when M is doubled, so that the scaled 28nm technology consumes much less space than the robust technology with the same architecture. In addition to shorter and thinner transistors in the scaled technology, the wirings, and especially the supply lines, can be narrower, because of the lower internal currents.

IV. TDCs AND TECHNOLOGY SCALING

The delay element presented in Fig. 2 can be utilized in the TDC structure presented in Fig. 3, for example. The rising edge of an external reference clock propagates into a delay line, which creates time samples for interpolation after every element (interpolation resolution = τ_p). The timing signals, start and stop, store the prevailing time sample in the register array, and the propagation delay of the delay-adjustable delay line is matched with the reference clock cycle time by means of a phase detector and charge pump. The timing diagram in Fig. 3 shows how the counter counts the full reference clock cycles between the timing signals and the interpolation (n=8) provides an accurate results for the time interval.

In the Nutt-based measurement architecture presented here the rms precision of the TDC, σ_{rms}, is compounded from several sources [3] and the rms quantization error σ_q depends on the interpolation resolution τ_p. The delay element propagation delay mismatch accumulates especially markedly in long delay lines and can be several LSBs (τ_p). $\sigma_{inl\text{-}st}$ and $\sigma_{inl\text{-}sp}$ are the standard deviations of these integral non-linearities in the start and stop interpolation channels. σ_{tdc} is the rms value of the random inherent TDC jitter, which benefits from good noise sensitivity. σ_{clk} is the rms reference clock jitter. When the error sources are not correlated their effects can be summed:

$$\sigma_{rms} = \sqrt{{\sigma_q}^2 + {\sigma_{inl-st}}^2 + {\sigma_{inl-sp}}^2 + {\sigma_{tdc}}^2 + {\sigma_{clk}}^2}\,. \quad (12)$$

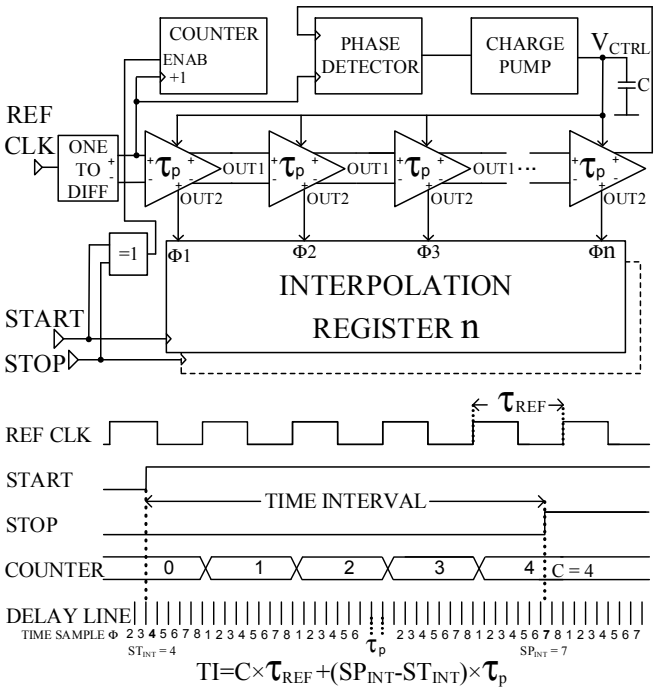

Fig. 3. A TDC using the proposed delay element.

One problem with the older CMOS technologies has been the quantization error, due to the large interpolation resolution τ_p. This has been resolved with a second interpolation level relying on sub-gate-delay interpolation methods [3], [5], [6]. σ_q in the architecture presented here is only a few picoseconds when modern technologies are used, which simplifies the TDC architecture significantly and reduces the size and power consumption.

The INL due to propagation delay mismatch nevertheless begins to dominate with modern technologies, as seen in (12), where the number of cascaded elements needed to be limited in order to restrict the accumulation of INL and a reference signal for the delay line then need to be created with an integrated high frequency oscillator, for example.

The jitter in the signals participating in the interpolation σ_{tdc} is expected to increase in modern technologies, so that more care needs to be taken when considering the layout and wiring. Signal slew rates need to be sufficient, especially in a TDC reaching picosecond level.

V. Conclusions

The technology comparison analysis performed here shows that digital cell propagation delay decreases by ~S/2 when technology is scaled down by a factor S. Smaller dimensions in transistors and shorter and narrower wirings, for example, lead to smaller load capacitances and higher speeds, and even the transistor intrinsic gain and supply voltage are smaller. In cascaded delay elements the propagation delay decreases down to a certain limit, at which the ratio W/L increases. This dramatically increases the area of the structure and the power consumption, however, so that oversizing is not good, either.

Digital cell propagation delay variation, when comparing the logic created in different process runs, $\sigma_{PROCESS}$, has improved in modern technologies, and this should be reflected in smaller variations in the maximum attainable processor clock frequency, for example. The delay variation within the circuit, $\sigma_{MISMATCH}$, however, does not scale together with the delay when modern technologies are used, and small imperfections in the processed layout do not decrease as much as does the transistor area. The smaller transistor dimensions together with weaker intrinsic gain emphasize small differences in transistor structure and create variations in the propagation delay. Increasing the transistor dimensions will reduce the role of the errors, but it will increase the size and power consumption.

Noise and other factors affecting the supply and control voltages have a greater influence on propagation delays in modern scaled technologies, and the smaller supply voltage also makes the active operating region of the transistor smaller and the relative effect of noise and other voltage error sources higher. On the other hand, power and area consumption are much less in modern scaled technologies.

The high speed logic used in the modern CMOS technology provides efficient blocks for processors and other digital signal processing, and the small propagation delays achieved in time domain signal processing can open the way to high-frequency oscillators and signal alternation with high resolution. TDCs with a resolution better than 10ps can be created with a simple gate-delay-based architecture, for example.

The propagation delay variation, $\sigma_{MISMATCH}$, does not usually have any critical influence on pure, high-speed digital systems, but its impact on time domain signal processing is higher. Parallel similar signal routes in clock trees or phase shifters, for example, can have more delay differences in modern technologies, and increases in the delay mismatch are especially harmful in picosecond-level TDCs, which are based on identical replicated delay generating structures and long chains of delay elements. The performance of TDCs does not scale with their resolution, because non-linearities, and also the increasing effects of noise, begin to weaken their performance. The older technologies give relatively small mismatch parameters, enabling precise TDCs to be fabricated as well. The use of "robust" CMOS technology can be reasonable especially in low volume productions where production costs are more important than size and power consumption.

References

[1] R. B. Staszewski, S. Vemulapalli, P. Vallur, J. Wallberg, and P. T. Balsara, "1.3 V 20 ps time-to-digital converter for frequency synthesis in 90-nm CMOS," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 53, no. 3, pp. 220–224, Mar. 2006.

[2] B. Murmann, "Digitally assisted analog circuits," *IEEE Micro*, vol. 26, no. 2, pp. 38–47, Mar./Apr. 2006.

[3] J.-P. Jansson, A. Mantyniemi, and J. Kostamovaara, "A CMOS time-to-digital converter with better than 10 ps single-shot-precision," *IEEE J. Solid-State Circuits*, vol. 41, no. 6, pp. 1286–1296, Jun. 2006.

[4] M. J. M. Pelgrom, A. C. J. Duinmaijer and A. P. G. Welbers, "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits*, vol. 24, no. 5, pp. 1433-1439, Oct 1989.

[5] K. Kim, W. Yu, and S. Cho, "A 9 bit, 1.12 ps resolution 2.5 b/stage pipelined time-to-digital converter in 65 nm CMOS using time-register," *IEEE J. Solid-State Circuits*, vol. 49, no. 4, pp. 1007–1016, Apr. 2014.

[6] P. Keränen and J. Kostamovaara, "A wide range, 4.2ps(rms) precision CMOS TDC with cyclic interpolators based on switched-frequency ring oscillators," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 62, no. 12, pp. 2795–2805, Dec. 2015.