2016 22nd IEEE International Symposium on Asynchronous Circuits and Systems

Analysis and Design of Delay Lines for Dynamic Voltage Scaling Applications

Ramy N. Tadros*, Weizhe Hua*, Matheus Gibiluka†, Matheus T. Moreira†, Ney L.V. Calazans†, and Peter A. Beerel*
*University of Southern California (USC) - Los Angeles, CA, United States
†Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS) - Porto Alegre, Brazil

***Abstract*—Dynamic voltage scaling of bundled-data asynchronous design has the promise to lead to far more energy-efficient systems than traditionally clocked alternatives. However, this approach relies on the development of energy-efficient delay lines, whose delay must track that of the combinational datapath over a wide range of voltages. This paper presents a thorough analysis of the design of such delay lines and describes how sizing affects their delay across different voltages. It proposes a design methodology for minimizing energy consumption subject to delay matching constraints. It then applies this methodology to delay lines that consist of four different delay elements in two different technologies, exploring the underlying trade-offs they present.**

I. Introduction

Asynchronous circuits' natural tolerance to variability and inherent flow control make them an attractive alternative to traditional synchronous designs in a wide-range of applications [1]. In particular, bundled-data asynchronous circuits have demonstrated significant benefits in both Network-on-Chips (e.g., [2]) and low-voltage compute blocks (e.g., [3], [4]). One challenge to these circuits is that they rely on delay lines (DLs) to control the synchronization of pipeline registers. These DLs must be energy efficient and provide a delay that is carefully matched to the worst-case-delay of the datapath under the expected range of process, voltage, and temperature (PVT) variations. These variations are particularly important in low and near-threshold voltages because the impact in delay is dramatically higher [5]. Moreover, there are a number of applications for which dynamic voltage scaling (DVS) is desired (e.g., [6], [7]), which implies that this matching constraint needs to be satisfied at a wide range of supply voltages.

Various DL designs exist in the literature to reliably delay a signal for a specific amount of time. These include tunable replica circuits (TRCs) [8], tunable delay buffers [9], and pre-charged inverter based DLs [10]. There have been some unorthodox approaches as well, such as using ring oscillators as the DL in [11]. Moreover, custom delay cells targeting programmability and fine-graining have also been proposed [12], [13]. Several of these works have focused on energy-efficient DLs and a few have explored their tolerance to PVT variations [8]. Only one of these, however, has explored designing DLs for voltage-scaling applications [14] and most perform transistor-sizing in an ad-hoc manner that may not apply when extending the design to a new technology.

Towards this goal, this paper develops an analytically-based methodology to design delay lines that have minimum energy subject to voltage-scaling-based matching constraints. In particular, this paper analyzes the design for supply voltages as low as near-threshold values. For lower voltages, the reader is directed to [15], [16] for related analyses.

The structure of the remainder of this paper is as follows. Section II provides background on DLs and a formal problem definition. Section III analyzes the factors affecting the problem and Section IV discusses our proposed design methodology. Subsequently, Section V presents our experimental results, including a methodology case study and a comparison between the different delay elements (DEs) architectures. Finally, Section VI provides some conclusions.

II. Problem Formulation

A. Delay Lines

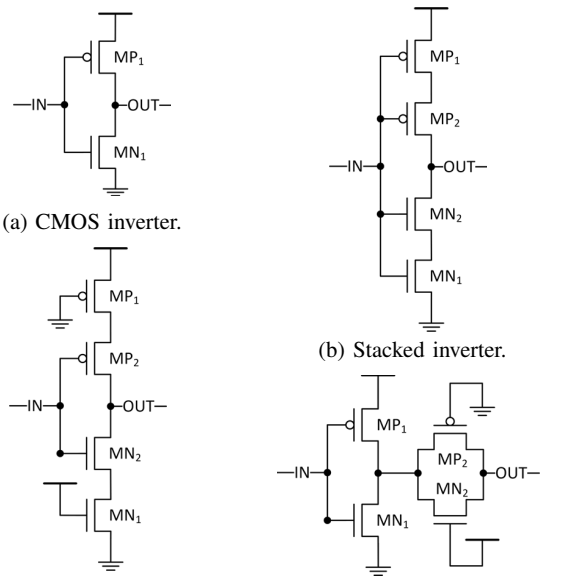
Many proposed DLs are formed by a chain of inverting CMOS gates. Four typical inverting gates, or DEs, are illustrated in Fig. 1. The first is simply the CMOS inverter (Fig. 1a). The second is the stacked CMOS inverter (Fig. 1b), where two MOS transistors in series are used to reduce the current while not increasing the diffusion output capacitance. This is different from reducing the current of the simple CMOS inverter because increasing transistor length affects diffusion capacitance and short channel effects affect the different DE structures differently. The third is a current starved inverter (CSI) (Fig. 1c), where the next-to-power-rails MOSs (MP_1 and MN_1) are kept always on. This decreases the current while not increasing either output diffusion or input gate capacitance. Lastly, Fig. 1d shows the use of a CMOS inverter and an added always-on transmission gate (TG) in series. This decreases current while not increasing input capacitance but does increase switched diffusion capacitance. One of the questions that this paper tries to answer is which one of the four building blocks should be used for constructing DLs and why, providing a guideline for designers.

Towards this end, Section III analyzes the CMOS inverter with the assumption that the fundamental findings and conclusions can be extended to other DEs architectures.

B. Problem Statement

Informally, we wish to choose a DL architecture and transistor sizing that achieves a target delay while minimizing energy consumption and matching the datapath across voltages. In particular, we assume that the design will not only operate at nominal voltage but at some arbitrary number of lower voltages as well via DVS. Our model of energy encompasses both switching and leakage energy assuming an average duty

978-1-4673-9007-1/16 $31.00 © 2016 IEEE
DOI 10.1109/ASYNC.2016.16



(a) CMOS inverter. (b) Stacked inverter. (c) Current starved inverter. (d) Inverter + TG.

Fig. 1: Different DEs architectures.

cycle at nominal voltage of the system period T_{sys}, similar to [15]. This point is detailed in Section III-C. We assume N different possible supply voltages and define a voltage specific activity factor r_i to indicate the percentage of time at which the system works at supply value i:

$$E = \sum_{i=1}^{N} r_i E_i \tag{1}$$

To quantify the notion of matching, we use the *voltage scaled delay ratio* (*VSDR*) introduced in [13], [14]:

$$\textit{VSDR} = \frac{T_{low}}{T_{nom}}, \tag{2}$$

where T_{low} and T_{nom} are the delay at a low and the nominal supply voltage, respectively. Without loss of generality, we assume the DL is designed to be larger than the delay of the combinational logic at the nominal supply voltage, considering both worst-case-delay and any PVT variations [17].[1] Then, for every other expected voltage, i, we introduce a two-sided *VSDR* constraint:

$$\textit{VSDR}_{functional_i} \leq \textit{VSDR}_i \leq \textit{VSDR}_{performance_i}, \tag{3}$$

The *VSDR* lower bound ($\textit{VSDR}_{functional_i}$) is a constant that guarantees that the DL slows down more than the combinational logic at the lower voltage. This guarantees at the lower voltage, the DL is still slower than the worst-case-delay of the combinational logic and all bundled-data constraints will still be met. The *VSDR* upper bound ($\textit{VSDR}_{performance_i}$) is a constant that bounds the unnecessary margin on the DL at lower voltages, and it should be set based on the system power budget. We thus formulate the problem definition as follows:

[1]In particular, assuming that the delay line was matched to the combinational logic at a lower supply voltage requires straight-forward changes to the equations that follow.

$$Min\big\{E(Sizing)\big\}, \qquad subject\,to \quad (3), \tag{4}$$

where E is the energy consumption of the DL as a function of its transistor sizing. In next sections, we analyze the relationship between transistor sizing, *VSDR*, and energy consumption.

III. Problem Analysis

A. Delay vs Sizing

To analyze how the MOS sizing affects the delay of a DL composed of a sequence of identical inverting gates, we return to the fundamentals of ICs design [18]–[20], where the gate delay, t, of every gate in the sequence can be modeled as:

$$t \propto \frac{C \cdot V_{DD}}{I}, \tag{5}$$

where C is the load capacitance seen by the gate, V_{DD} is the supply voltage, and I is the average current flowing into the capacitance during this period of time t. Based on this simple model and the understanding of the logical effort [19], it is often assumed the delay is not affected by changing the width of all transistors, ignoring the effect of the interconnects delay. In particular, increasing the width of the transistors in each inverting gate increases their current but also increases the load, cancelling each other out. However, this section shows that when independently sizing pull-up and pull-down networks, this is not always the case.

We know that the sizing does not affect V_{DD}. Regarding the capacitance, it is well established that at any supply value:

$$C \propto (W_p L_p + W_n L_n), \tag{6}$$

where W_p, L_p, W_n, L_n are the width and length of the pMOS and nMOS respectively. Also, since our focus is on a sequence of identical inverting gates, it is irrelevant to either separate the diffusion and gate capacitances, or the input and the output gates. Regarding the current, for super-threshold operations, we adopt the alpha-power law model [21], where the current depends on the following:

$$I \propto \mu \frac{W}{L} \big(V_{GS} - V_{th}\big)^{\alpha}, \tag{7}$$

where μ is the carrier mobility, W and L are the MOS width and length, V_{GS} can be considered equal to V_{DD} throughout our analysis, V_{th} is the threshold voltage, and α is the velocity saturation index, which is a technology dependent empirical coefficient. The delay of the DL is composed of two components: rise and fall. Then for an even number of identical inverting gates, the total delay can be written:

$$T \propto t_{rise} + t_{fall}, \tag{8}$$

where t_{rise} is related to the pull-up network, i.e., the pMOSes, and t_{fall} to the nMOSes. Then substituting (5), (6), and (7) into (8). We neglect the change in threshold with the sizing for simplicity in this subsection, which is an acceptable approximation at nominal supply value, then all the voltages cancel each other off, and the delay at nominal supply is:

$$T \propto \frac{(W_p L_p + W_n L_n)}{\mu_p W_p / L_p} + \frac{(W_p L_p + W_n L_n)}{\mu_n W_n / L_n}, \tag{9}$$

In this subsection, we fix L_n and L_p to be equal values and study the resulting relation between delay and width:

$$T \propto (W_p + W_n) \cdot \left(\frac{1}{\mu_p W_p} + \frac{1}{\mu_n W_n} \right), \tag{10}$$

In this way, T can be approximated differently depending on the relation between $\mu_p W_p$ and $\mu_n W_n$. In case (i) $\mu_n W_n \gg \mu_p W_p$, the term $1/\mu_n W_n$ can be neglected. In case (ii) $\mu_n W_n \sim \mu_p W_p$, we can assume that $\mu_p W_p = \mu_n W_n$. Finally, in case (iii) $\mu_n W_n \ll \mu_p W_p$, the term $1/\mu_p W_p$ can be neglected. This leads us to the following form of (10):

$$T \propto \begin{cases} (i)\ \mu_n W_n \gg \mu_p W_p \Rightarrow \dfrac{1}{\mu_p}\left(\dfrac{W_n}{W_p} + 1\right) \\ (ii)\ \mu_n W_n \sim \mu_p W_p \Rightarrow 2\left(\dfrac{1}{\mu_p} + \dfrac{1}{\mu_n}\right) = const \\ (iii)\ \mu_n W_n \ll \mu_p W_p \Rightarrow \dfrac{1}{\mu_n}\left(\dfrac{W_p}{W_n} + 1\right) \end{cases} \tag{11}$$

Thus, sizing does not impact delay when the relative strengths of the pull-up and pull-down networks are approximately the same. Otherwise, sizing does affect the delay.

Fig. 2 shows the simulation results of the delay of a sequence of ten CMOS inverters with minimum length while varying the width as shown. The results are shown in two different technologies: a 65nm bulk CMOS and a 28nm UTBB FDSOI, both at nominal supply equal to 1V. This set of curves shows that the analysis presented in (11) explains the delay behavior subject to sizing. For example, the blue curve in Figs. 2b and 2d starts in case (i) where the delay is inversely proportional to W_p, then ends in case (iii) where it is directly proportional, in a parabolic-like behavior. While the black curve in Figs. 2a and 2c starts in case (iii) then enters case (ii) and never reaches case (iii) due to the large W_p. The understanding of these observations are useful when designing DEs for optimum energy, as discussed in Section IV.

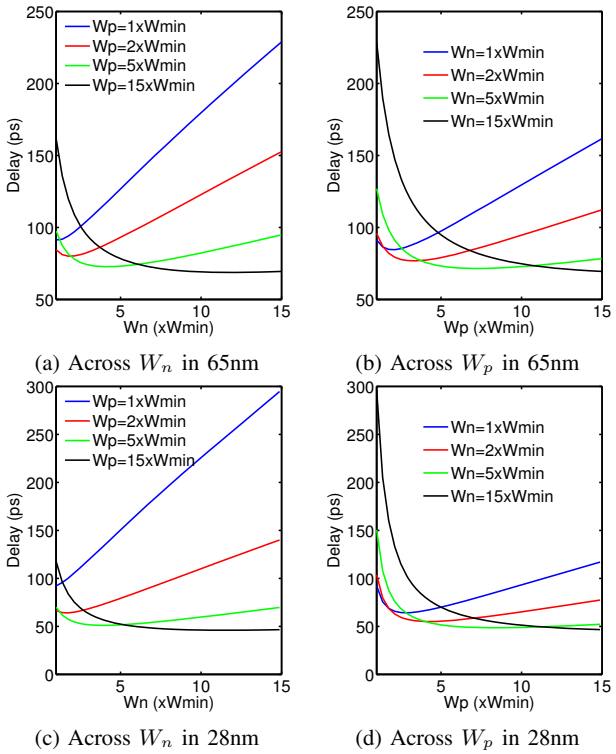
B. VSDR Dependencies

In this subsection, we try to explain the behavior of the delay compared across different supply voltages. The metric *VSDR* is defined in (2) to quantify that comparison. First, we quantify the delay of a sequence of identical inverters following (5) as follows:

$$T \propto \frac{C_{load} \cdot V_{DD}}{I_{avg}}, \tag{12}$$

where C_{load} is the total load capacitance of the DL, I_{avg} is the average current drawn from the supply during this period of time T ignoring leakage, and the average between rise and fall is considered. We substitute (12) and (7) into (2). For simplicity, we ignore the differences between nMOS and pMOS in α and μ, and only consider a single average, which is acceptable for an approximate analysis. We get

$$VSDR \propto \frac{C_{loadlow} \cdot V_{DDlow}}{C_{loadnom} \cdot V_{DDnom}} \cdot \frac{\left(V_{GSnom} - V_{thnom}\right)^{\alpha}}{\left(V_{GSlow} - V_{thlow}\right)^{\alpha}}, \tag{13}$$



(a) Across W_n in 65nm (b) Across W_p in 65nm

(c) Across W_n in 28nm (d) Across W_p in 28nm

Fig. 2: The average delay of a sequence of ten identical inverters. All lengths are equal to minimum length (*Lmin*). All widths are scaled to minimum width (*Wmin*).

The supply ratio is a constant so it can be removed from the relation. The critical factors in this equation are the threshold voltages which are affected by the transistor sizes differently depending on a variety of second order effects. These include short channel effects (SCE), narrow channel effects (NCE), reverse SCE, and reverse NCE [22]–[24]. Because some of these effects have different trends and correlations with length and width depending on the technology, we evaluate our designs in two disparate technologies, a bulk CMOS 65nm technology and an FDSOI 28nm technology. Even for SOI, which uses a PSP model instead of BSIM, these second order effects persist. Based on [18], [19], [22]–[28], and for simplicity, we decompose the V_{th} as follows:

$$V_{th} = V_{th0} - \Delta V_{th}(W, L, V_{DD}), \tag{14}$$

where V_{th0} is the threshold voltage at infinite size and no body effect. This encompasses all the various effects into a single term. The dependency on the supply voltage is due to the Drain Induced Barrier Lowering (DIBL) effect [23], [24]. which can be either positive or negative, depending on the technology. We decompose the overdrive voltage as follows:

$$V_{ov} = V_{GS} - V_{th} = V_{GS} - V_{th0} + \Delta V_{th} = OV + \Delta V_{th}, \tag{15}$$

Now, (13) can be re-written:

$$VSDR \propto \frac{C_{loadlow}}{C_{loadnom}} \cdot \left(\frac{OV_{nom} + \Delta V_{thnom}}{OV_{low} + \Delta V_{thlow}} \right)^{\alpha}, \tag{16}$$

It is safe to neglect ΔV_{thnom} compared to the value of OV_{nom}, however, as V_{DD} gets lower and closer to the V_{th} value, ΔV_{thlow} becomes more substantial with respect to the value of OV_{low}, hence it cannot be neglected. Dividing (16) by OV_{low}, we get

$$VSDR \propto \frac{C_{loadlow}}{C_{loadnom}} \cdot \left(\frac{OV_{nom}}{OV_{low}}\right)^{\alpha} \cdot \frac{1}{\left(1+\frac{\Delta V_{thlow}}{OV_{low}}\right)^{\alpha}}, \quad (17)$$

The OV ratio is a constant which can be removed from the relation. For a supply voltage sufficiently higher than V_{th}, it is safe to say that the term ΔV_{thlow} is smaller than OV_{low}, and hence a binomial approximation can be applied:

$$VSDR \propto \frac{C_{loadlow}}{C_{loadnom}} \cdot \left(1-\alpha\frac{\Delta V_{thlow}}{OV_{low}}\right), \quad (18)$$

Because α and OV_{low} are constants relative to sizing, a closed form for the *VSDR* can be derived, and curve fitting can be used to verify the following final equation:

$$\boxed{VSDR = \frac{C_{loadlow}}{C_{loadnom}} \cdot \left(k_1 - k_2 \Delta V_{thlow}\right),} \quad (19)$$

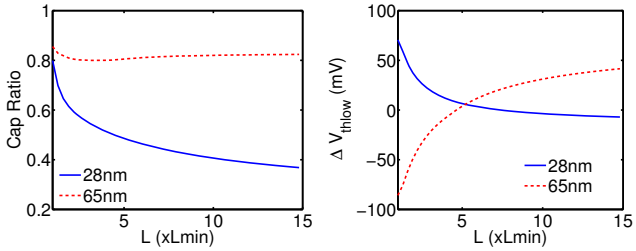
where k_1 and k_2 are fitting parameters, both of which have to be positive as per our analysis.

Equation (19) tells us that the *VSDR* depends on two factors, both depending on the threshold voltage. All experimental results use $V_{low} = 0.6V$.

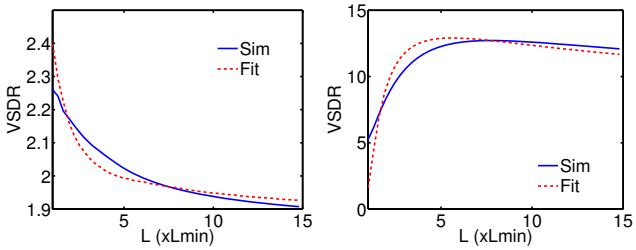
i) $C_{loadlow}/C_{loadnom}$: From [29], we know that this ratio is smaller than one due to lower C_{ox} resulting from the larger V_{th}/V_{DD} ratio. However, due to the complexity of analyzing this capacitance ratio, we will deal with it as a measured quantity. The capacitances were measured in the same inverters sequence setup used in Section III-A. They were measured in SPICE using the current integration method, averaged over rise and fall transitions. Figs. 3a and 4a show this capacitance ratio versus length and width respectively.

ii) ΔV_{thlow}: which is the threshold variation at low supply voltage. As previously discussed, this amount is strongly dependent on the technology. They were measured in SPICE using the evaluated sub-circuit MOS threshold, averaged over nMOS and pMOS. Figs. 3b and 4b show the ΔV_{thlow} versus length and width respectively.

In order to verify the correctness of the previous analysis, we curve fitted equation (19) to the measured results in both 28nm and 65nm versus both length and width. For the sequence of ten identical inverters, three quantities were measured: *VSDR*, capacitance ratio, and ΔV_{thlow}. *VSDR* is divided by capacitance ratio, then the output is fitted to the relation in (19), using the MATLAB *lsqcurvefit* function, with the *trust region reflective* algorithm. Then the right hand side is evaluated by multiplying the fitted function by the capacitance ratio. Both the measured value of the *VSDR* and the fitted one are plotted versus length using point-to-point mapping, the same as the mapping used for the capacitance ratio and

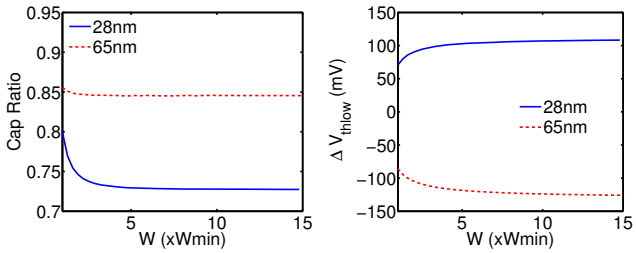


(a) Capacitance Ratio (b) ΔV_{thlow}

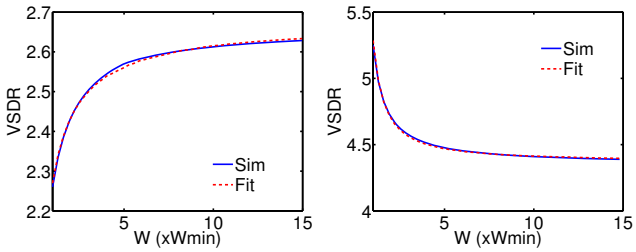


(c) *VSDR* in 65nm (d) *VSDR* in 28nm

Fig. 3: *VSDR* fitting curves across L. W_n=$W_p/2$=*Wmin*.



(a) Capacitance Ratio (b) ΔV_{thlow}



(c) *VSDR* in 65nm (d) *VSDR* in 28nm

Fig. 4: *VSDR* fitting curves across W. L_n=L_p=*Lmin*.

ΔV_{thlow}. The fitted plots, parameters, and mean square errors are as follows: (i) across length in 65nm is in Fig. 3c with k_1=2.49, k_2=3.66, R^2=0.07, (ii) across length in 28nm is in Fig. 3d with k_1=29.00, k_2=382.19, R^2=84.84, (iii) across width in 65nm is in Fig. 4c with k_1=1.66, k_2=11.54, R^2=0.00, (iv) across width in 28nm is in Fig. 4d with k_1=7.68, k_2=15.09, R^2=0.01. These results support our assumptions and analysis that sizing does alter *VSDR* due to primarily capacitance ratio and ΔV_{thlow} and that the trends strongly depends on the choice of technology. They also show how equation (19) successfully fits the *VSDR* even in disparate technologies.

C. Energy Efficiency

The last piece of the puzzle, formulated in (4), is the energy consumption. We start by the average power, which has two components:

$$P = P_{static} + P_{dynamic}, \tag{20}$$

where P_{static} is the leakage power [15]:

$$P_{static} = V_{DD} I_{off}, \tag{21}$$

where I_{off} is the sub-threshold current (leakage current), which equals [15], [22], [23]:

$$I_{off} = I_0 \frac{W}{L} e^{(V_{GS}-V_{th})/n \cdot v_t} \left(1 - e^{-V_{DS}/v_t}\right) \tag{22}$$

where I_0 is the technology dependent sub-threshold current extrapolated for $V_{GS} = V_{th}$, n is the sub-threshold factor, v_t is the thermal voltage, and V_{DS} is the drain source voltage.

$P_{dynamic}$ only depends on the energy per transition (EPT) ignoring the short circuit power [15] and it is well established that [20]:

$$EPT = C_{load} V_{DD}^2, \tag{23}$$

where C_{load} is the total capacitance switched during one transition through the DL. However, the combination of both static and dynamic components depends on the relative activity of the DL. Since T is typically much smaller than the global system period T_{sys}, i.e., the DL is active for a certain duty cycle of T_{sys}, as defined in Section II. Re-formulating the analysis in [15] to conform with our application, (20) becomes:

$$P = P_{static} + \frac{EPT}{T_{sys}}, \tag{24}$$

In order to get the average energy, E, over the system, P needs to be multiplied by T_{sys}. Substituting (21) and (23) into (24), we get

$$E = P \cdot T_{sys} = V_{DD} I_{off} T_{sys} + C_{load} V_{DD}^2. \tag{25}$$

Then using the proportionality in (12), we get

$$\boxed{E = V_{DD}\left(I_{off} T_{sys} + k I_{avg} T\right),} \tag{26}$$

where k is a proportionality constant.

IV. Proposed Sizing Methodology

Our proposed design methodology is to optimize the design of a sequence of identical inverting gates and scale this sequence to achieve the target delay. Towards this end, rather than directly minimizing energy of a DL with a target delay, we propose the equivalent strategy of minimizing *energy-per-delay* (E/T) of the DL. Dividing (26) by T we obtain:

$$E/T = V_{DD} \left(\frac{I_{off}}{\beta} + k I_{avg}\right), \tag{27}$$

where β is the activity duty cycle of the optimized DL from the system point view, i.e., T/T_{sys}. Therefore, E/T depends solely on the current. Hence, based on (7) and (22), E/T is expected to be largely proportional to the transistor width and inversely proportional to the length.

Based on this analysis, and from the analysis of Section III-A, the formulation stated in (4) can be re-written as follows:

$$\boxed{Min\Big\{ E/T\left(W_p, W_n, L_p, L_n\right)\Big\}, \quad subject\,to\,(3),} \tag{28}$$

where E/T depends on the current according to (27) and is weighted according to the DVS system specifications in (1).

Since E/T is an increasing function of W, and a non-increasing function of L, the smallest possible W and largest possible L should be always selected as long as the *VSDR* constraints are satisfied. Interestingly, based on the *VSDR* analysis in Section III (see Figs. 3 and 4), the *VSDR* showed inconsistent correlations with L and W at different technologies. This leads to a 4-dimensional design space (W_p, W_n, L_p, L_n) and a complete case study is presented in Section V-B. To better appreciate practical approaches for finding the optimal solution in this design space, this section describes a solution for two possible scenarios in which we neglect the difference between nMOS and pMOS for simplicity, reducing the design space to 2 dimensions.

Scenario I - VSDR is a decreasing function in L (As in the 65nm case): Since increasing length results in less energy, then at any sizing, increasing L to $L + dL$ will always result in lower energy where the *VSDR* is the only limiter, and hence $VSDR_{functional}$ becomes the bottleneck. Thus, the most energy efficient sizing is found at the lower bound of the *VSDR* constraint, which is a fact independent of W. Two sizing strategies can be used depending on the following: (i) If *VSDR* is a non-increasing function in W, then decreasing W becomes unbounded by neither the *VSDR* constraint nor the energy, and we use the technology hard bound *Wmin* while increasing L until hitting the value of $VSDR_{functional}$. (ii) If *VSDR* is a non-decreasing function in W, then we have the same argument as L, decreasing W to $W - dW$ will always result in lower energy, and the most energy efficient point is met at the bottleneck $VSDR_{functional}$, which is a fact independent of L. Then we do a brute force search of all the combinations of L and W that result in $VSDR_{functional}$. Compare E/T of these combinations and select the sizing that achieves the minimum E/T.

Scenario II - VSDR is an increasing function in L (As in the 28nm case): Similarly, since increasing length results in less energy, $VSDR_{performance}$ becomes the bottleneck. Thus, the most energy efficient sizing is found at the upper bound of the *VSDR* constraint. Then, we have two cases as well: (i) If *VSDR* is a non-increasing function in W, then we do a brute force search of all the combinations of L and W that result in $VSDR_{performance}$. Compare their E/T and choose the one that achieves the minimum E/T. (ii) If *VSDR* is a non-decreasing function in W, then use the minimum W while increasing L until hitting the value of $VSDR_{performance}$.

It is worth mentioning that for applications that care the most about the energy efficiency, a technology that follows *Scenario II* is a better fit because it gives the designer an arbitrary bound on performance which he can trade off for

more energy savings. A technology following *Scenario I*, on the other hand, is limited by the functionality and further energy savings will result in a timing violation.

It is also important to note that the feasibility of the solution is not always guaranteed. This happens when no feasible sizing would make the $VSDR_{DL}$ satisfies the constraints, leaving the designer with two options: use a different DE as discussed in Section V-D, or trade off the performance in one of two ways according to the scenario: using a longer DL which will lower the performance at nominal supply in the case when the *VSDR* is too low (*Scenario I*) or increasing the performance bound when *VSDR* is too high (*Scenario II*).

V. Experimental Results

The objective of this paper is to size the DLs to minimize E/T subject to *VSDR* constraints and compare the four DEs illustrated in Fig. 1 after optimization. Towards this goal, this section first simulates the critical paths of ISCAS'85 benchmarks to motivate the *VSDR* constraints for our comparison. Because the results are highly process dependent, we performed this analysis on both bulk 65nm and SOI 28nm. Moreover, to improve the clarity of the approach, we include a detailed case study of our optimal sizing procedure for a DL in the 28nm process.

A. VSDR Specifications

As previously discussed, the lower bound of the *VSDR* constraint in (3) is tied to the *VSDR* characteristics of the combinational logic matched by the DL. We synthesized each ISCAS benchmark to the target technology using Synopsys Design Compiler. Using Synopsys PrimeTime, we then extracted a SPICE deck describing its critical path. We then measured the delay of the critical path at each of the target voltages. Then the *VSDR* is computed using (2). The simulation results are summarized in Table I with $V_{low} = 0.6V$.

B. Methodology Case Study

To illustrate the sizing methodology proposed in Section IV, this subsection applies the methodology to an inverter based DL that is intended to match the combinational logic of the critical path of the *c2670* benchmark in our 28nm technology. Assuming a single value of $V_{low} = 0.6V$, with $r_{nom} = r_{low} = 0.5$ in (1), a fixed value of T_{sys} arbitrary set to an order of magnitude higher than the range of T, and $k = 1$ for simplicity in (27). The optimization is done for a sequence of 20 identical inverters. First, it is important to figure out the correlation between W_p, L_p, W_n, L_n and *VSDR*. According to our experiments, the *VSDR* is a decreasing function in W_p and an increasing function in the other three, which fits into *Scenario II*. As previously explained, the strategy is hence to adopt minimum W_n and trade off between L_n, L_p, and W_p. From Table I, *VSDR* of benchmark *c2670* is 5.40, which is $VSDR_{functional}$ in (3). Assuming a 10% allowed performance loss, $VSDR_{performance}$ equals 5.94. Then, the problem is reduced to a 3-D search which can be elaborated specifically for this case as follows:

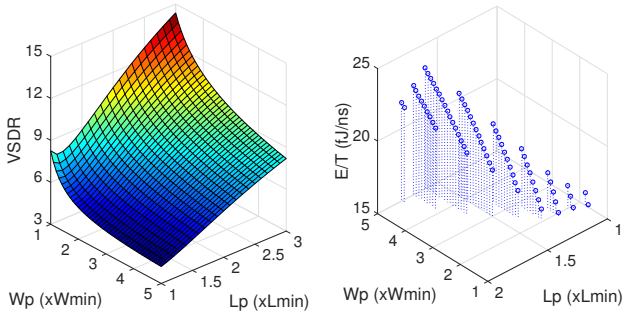


Fig. 5: One iteration of the $while\ loop$ at $L_n = Lmin$

1: $L_n = Lmin$
2: **while** $L_n < 3Lmin$ **do**
3: Measure *VSDR* versus L_p and W_p.
4: Find the combinations satisfying the target *VSDR* within a reasonable tolerance.
5: Measure the E/T at these combinations.
6: Record the combination achieving minimum E/T at that corresponding L_n.
7: $L_n = L_n + 0.1Lmin$
8: **end while**
9: Plot the obtained E/T versus L_n.
10: Use the sizing corresponding to the min point of the plot.

Fig. 5 shows one iteration of the $while\ loop$ as an example at a certain value of L_n. It shows a 3D-plot of *VSDR* versus W_p and L_p. The combinations satisfying the interval are selected, and their E/T values are also plotted. And hence, the combination achieving minimum E/T at that L_n is found and recorded. Then the best sizing is found across L_n values, which is found to be $(L_n, L_p, W_n, W_p) = (1.2X, 1X, 1X, 1.6X)$ in terms of the minimum sizes, achieving the best E/T of 13.42fJ/ns with a *VSDR* = 5.93.

C. Extended VSDR Analysis

This subsection extends the analysis of *VSDR* on the inverter based DL in Section III-B to the *VSDR* of the other three architectures of Fig. 1. From (27) we concluded that E/T only depends on the current, but the optimal structure depends on which provides the least current while also considering the *VSDR* constraint. According to (19), *VSDR* depends on the load capacitance ratio ($C_{load_{low}}/C_{load_{nom}}$) and the threshold variation at the low supply voltage ($\Delta V_{th_{low}}$). To better evaluate this constraint we studied the behavior of these quantities separately in both the 28nm and 65nm processes. All the transistors are sized the same to be able to distinct the architectural differences only. The study is done as a function of transistor length but similar trends exist for width.

We first measured and plotted the capacitance ratio across L in Fig. 6. From the fundamentals of the different MOS capacitances discussed in [18], [19], we know that the diffusion capacitances are more voltage dependent than their gate counterparts. Therefore, it is expected that the capacitance ratio will depend on the diffusion to gate ratio. In other

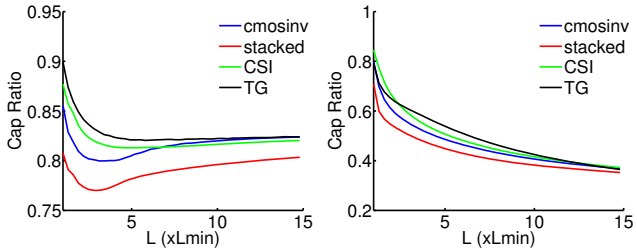
TABLE I: *VSDR* of ISCAS'85 benchmark evaluated at 600mV

Tech	*c17*	*c432*	*c499*	*c880a*	*c1355*	*c1908*	*c1908a*	*c2670*	*c2670a*	*c3540*	*c3540a*	*c5315*	*c5315a*	*c6288*	*c7552*
28nm	5.70	5.38	5.99	5.42	5.87	5.53	5.41	5.40	5.37	5.31	5.42	5.38	5.79	6.19	5.86
65nm	2.76	2.64	2.72	2.74	2.70	2.70	2.73	2.69	2.76	2.61	2.63	2.71	2.71	2.74	2.77

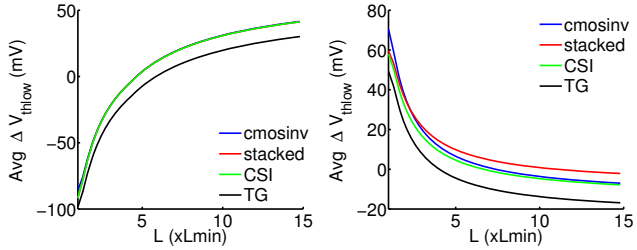
words, the architecture with relatively higher diffusion to gate capacitance, should have higher capacitance ratio. The TG-based has four diffusion capacitances seen from the output node, along with the same gate capacitance as the CMOS inverter, hence it should have the highest capacitance ratio. The stacked has double the gate capacitance but the same diffusion, hence it should have the lowest capacitance ratio. The CSI has additional internal capacitance that is translated to the output node through the on MOS, then it should have higher capacitance ratio than the CMOS inverter. Our experimental results shown in Figs. 6a and 6b validate this analysis.

We secondly measured and plotted the threshold variation across L in Fig. 6. The average threshold over all transistors is reported. One difference between them is caused by a second order effect dependent on the V_{DS} of the transistors. From [22]–[24], we know that DIBL is a non-decreasing function in V_{DS} and that the larger the V_{DS} across a MOS, the larger ΔV_{th}. For the TG-based, the TG is a non-inverting gate, its transistors have the lowest V_{DS}, and hence the lowest ΔV_{th}. In contrast, the CMOS inverter has the largest V_{DS} and hence the highest ΔV_{th}. For the CSI and the stacked DEs, it is averaged over two transistors, but MP_1 and MN_1 in CSI have much lower V_{DS} across them since their drains act as virtual power rails. This is verified in Figs. 6c and 6d for relatively low lengths. The other difference is caused by the body effect [18], [19], which is severe in the stacked one, causing ΔV_{th} to decrease. The body effect coefficient decreases with L in SOI technologies [25] opposite to old trends [30]. Also, DIBL decreases with L in any technology, but it causes ΔV_{th} to increase, similar to many other SCE. Consequently, as L increases, the stacked ΔV_{th} exceeds that of the CMOS inverter. Note that body effects are higher in SOI technologies [31], that is why the curves show these effects more clearly in 28nm SOI than in 65nm bulk.

In summary, some of the trends for the *VSDR* constraints in (3) strongly depend on the technology while others do not. For example, the stacking will suffer from body effect, which will increase the *VSDR*, however, it has the lowest capacitance ratio, which will decrease the *VSDR*. The body effect dominates in SOI 28nm while the capacitance ratio does in bulk 65nm, which affects the optimal DE architecture. On the other hand, the TG is expected to give the highest *VSDR* at same sizing, due to the largest capacitance ratio and lowest ΔV_{th}, independent of technology. Moreover, it has an additional MOS assisting the load switching, hence it will be faster and thus have relatively higher E/T at the same sizing and current value. Similarly, while the CSI seems efficient because it limits the current while not affecting the switching capacitance, (27) showed that this benefit does not help E/T independent of technology.



(a) Capacitance Ratio in 65nm (b) Capacitance Ratio in 28nm



(c) ΔV_{thlow} in 65nm (d) ΔV_{thlow} in 28nm

Fig. 6: Capacitance ratio and ΔV_{thlow} across L. W_n=$W_p/2$=*Wmin*.

D. Final Comparison

This subsection extends the case study presented in Section V-B to all DE architectures in both technologies. The DEs lengths and widths were sized targeting optimum E/T for a short sequence of identical DEs and then extended to a fixed T of 1.2ns in 28nm and 1.6ns in 65nm. The results are summarized in Table II. T_{sys} is set to $10T$ for both technologies. Monte-Carlo simulation with 1000 samples modeling local mismatches at the three extreme corners for PVT are shown in the *MC VSDR* field, where the *VSDR* mean value (μ) and standard deviation (σ) are reported. The area values shown are active area only, i.e. $\sum_i W_i L_i$.

As 65nm follows *Scenario I* in which larger L reduces *VSDR* and the optimum sizing is found at the lower bound of *VSDR*, the optimum E/T is achieved by the TG-based DE because its relatively high *VSDR* enables larger L and thus lower E/T. On the other hand, 28nm follows Scenario II in which the optimal point is found at the upper bound of *VSDR*. Here, the optimum E/T is achieved by the stacked structure which provides the lowest relative *VSDR* due to a trade-off between capacitance ratio and threshold variation (as discussed in Section V-C). Interestingly, these results show that when the sizing is constrained to satisfy the VSDR constraints in (3) the inverter based DE in 28nm is actually larger than the stacked DE. Finally, note that the CMOS inverter in 65nm is an example of a non-existent feasible solution to the problem (see Section IV) and can only be used at the expense of lower performance at nominal supply.

TABLE II: Comparison Summary

				Inverter	CSI	Stacked	TG-Inv
65 nm	*E/T (fJ/ns)*			N/A	62.78	45.85	31.53
	MC VSDR	*slow*	μ	N/A	2.89	2.93	2.87
			σ	N/A	0.64%	0.85%	1.19%
		typ	μ	N/A	2.70	2.70	2.73
			σ	N/A	0.58%	0.71%	1.08%
		fast	μ	N/A	2.28	2.28	2.33
			σ	N/A	0.43%	0.47%	0.95%
	Area (μm^2)			N/A	25.91	7.63	4.59
28 nm	*E/T (fJ/ns)*			13.88	9.45	7.69	12.72
	MC VSDR	*slow*	μ	19.38	17.70	17.80	17.48
			σ	70.13%	45.30%	66.42%	46.90%
		typ	μ	5.94	5.94	5.94	5.94
			σ	8.47%	6.12%	8.77%	6.85%
		fast	μ	2.90	3.00	3.04	3.00
			σ	1.76%	1.54%	2.04%	1.71%
	Area (μm^2)			1.05	1.50	0.77	1.26

VI. Conclusion

This paper presents a thorough analysis of the design of delay lines for voltage scaling applications. It discusses how sizing affects the delay of a DL, it demystifies the relation between the sizing and the relative delay of the DL at lower voltages, and explains how a DL's energy efficiency should be quantified and compared. Based on this analysis, the paper proposes a design methodology targeting minimum energy consumption while maintaining delay matching requirements imposed by system specifications. The paper proves how that the optimal structure is a strong function of technology and second order effects. Also, it discusses the differences between possible architectures of delay elements and compare them in two different technologies. We applied this methodology to an arbitrary system specification and showed that the TG-based and the stacked structures achieve the optimum energy in 65nm and 28nm, respectively. This methodology can be used to design a library of efficient delay elements targeting a range of different delay matching constraints.

References

[1] M. Singh and S. Nowick, "High-throughput asynchronous pipelines for fine-grain dynamic datapaths," in *ASYNC*, 2000, pp. 198–209.

[2] A. Ghiribaldi, D. Bertozzi, and S. M. Nowick, "A transition-signaling bundled data NoC switch architecture for cost-effective GALS multicore systems," in *DATE*, Mar 2013, pp. 332–337.

[3] A. Sridharan, C. Sechen, and R. Jafari, "Low-voltage low-overhead asynchronous logic," in *ISLPED*, Sept 2013, pp. 261–266.

[4] I. J. Chang, S. P. Park, and K. Roy, "Exploring asynchronous design techniques for process-tolerant and energy-efficient subthreshold operation," *IEEE JSSC*, vol. 45, no. 2, pp. 401–410, Feb 2010.

[5] S. Hanson, B. Zhai, K. Bernstein, D. Blaauw, A. Bryant, L. Chang, K. K. Das, W. Haensch, E. J. Nowak, and D. M. Sylvester, "Ultralow-voltage, minimum-energy CMOS," *IBM J. Res. Dev.*, vol. 50, no. 4.5, pp. 469–490, Jul 2006.

[6] Y. Pu, J. Pineda de Gyvez, H. Corporaal, and Y. Ha, "An ultra-low-energy multi-standard JPEG co-processor in 65 nm CMOS with sub/near threshold supply voltage," *IEEE JSSC*, vol. 45, no. 3, pp. 668–680, Mar 2010.

[7] Z. Wang, X. He, and C. Sechen, "TonyChopper: A desynchronization package," in *ICCAD*, Nov 2014, pp. 446–453.

[8] J. Tschanz, K. Bowman, S. Walstra, M. Agostinelli, T. Karnik, and V. De, "Tunable replica circuits and adaptive voltage-frequency techniques for dynamic voltage, temperature, and aging variation tolerance," in *VLSIC*, Jun 2009, pp. 112–113.

[9] A. Chakraborty, K. Duraisami, A. Sathanur, P. Sithambaram, L. Benini, A. Macii, E. Macii, and M. Poncino, "Dynamic thermal clock skew compensation using tunable delay buffers," *IEEE TVLSI*, vol. 16, no. 6, pp. 639–649, Jun 2008.

[10] G. Li, Y. Tousi, A. Hassibi, and E. Afshari, "Delay-line-based analog-to-digital converters," *IEEE TCAS-II*, vol. 56, no. 6, pp. 464–468, Jun 2009.

[11] P. Lu, A. Liscidini, and P. Andreani, "A 3.6 mw, 90 nm CMOS gated-vernier time-to-digital converter with an equivalent resolution of 3.2 ps," *IEEE JSSC*, vol. 47, no. 7, pp. 1626–1635, Jul 2012.

[12] M. Maymandi-Nejad and M. Sachdev, "A digitally programmable delay element: design and analysis," *IEEE TVLSI*, vol. 11, no. 5, pp. 871–878, Oct 2003.

[13] A. Singhvi, M. T. Moreira, R. N. Tadros, N. L. Calazans, and P. A. Beerel, "A fine-grained, uniform, energy-efficient delay element for FD-SOI technologies," in *ISVLSI*, Jul 2015, pp. 27–32.

[14] G. Heck, L. Heck, A. Singhvi, M. Moreira, P. Beerel, and N. Calazans, "Analysis and optimization of programmable delay elements for 2-phase bundled-data circuits," in *VLSID*, Jan 2015, pp. 321–326.

[15] M. Alioto, "Ultra-low power VLSI circuit design demystified and explained: A tutorial," *IEEE TCAS-I*, vol. 59, no. 1, pp. 3–29, Jan 2012.

[16] D. Markovic, C. Wang, L. Alarcon, T.-T. Liu, and J. Rabaey, "Ultralow-power design in near-threshold region," *Proc. IEEE*, vol. 98, no. 2, pp. 237–252, Feb 2010.

[17] I. Chang, S. Park, and K. Roy, "Exploring asynchronous design techniques for process-tolerant and energy-efficient subthreshold operation," *IEEE JSSC*, vol. 45, no. 2, pp. 401–410, 2010.

[18] S. Kang and Y. Leblebici, *CMOS Digital Integrated Circuits*. McGraw-Hill, 2003.

[19] N. Weste and D. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*. Addison Wesley Publishing Company Incorporated, 2011.

[20] A. Chandrakasan, S. Sheng, and R. Brodersen, "Low-power CMOS digital design," *IEEE JSSC*, vol. 27, no. 4, pp. 473–484, Apr 1992.

[21] T. Sakurai and A. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE JSSC*, vol. 25, no. 2, pp. 584–594, Apr 1990.

[22] Y. Tsividis, *Operation and Modeling of the MOS Transistor*. Oxford University Press, 1999.

[23] Y. Cheng and C. Hu, *MOSFET Modeling & BSIM3 User's Guide*. Kluwer Academic Publishers, 2002.

[24] W. Liu and C. Hu, *BSIM4 and MOSFET Modeling for IC Simulation*, ser. International series on advances in solid state electronics and technology. World Scientific, 2011.

[25] W. Yang, C.-H. Lin, T. H. Morshed, D. Lu, A. Niknejad, and C. Hu, *BSIMSOIv4.4 MOSFET model users' manual*. The Regents of the University of California, 2010.

[26] O. Rozeau, M.-A. Jaud, T. Poiroux, and M. Benosman, *UTSOI Model 1.1.3*. Laboratoire d'électronique et de technologie de l'information (Leti), May 2012.

[27] T. Chen and T. Gildenblat, "Analytical approximation for the MOSFET surface potential," *Solid-State Electronics*, vol. 45, no. 2, pp. 335 – 339, 2001.

[28] G. Gildenblat, X. Li, W. Wu, H. Wang, A. Jha, R. van Langevelde, G. Smit, A. Scholten, and D. Klaassen, "PSP: An advanced surface-potential-based MOSFET model for circuit simulation," *IEEE Trans. Electron Devices*, vol. 53, no. 9, pp. 1979–1993, Sept 2006.

[29] K. Nose, S.-I. Chae, and T. Sakurai, "Voltage dependent gate capacitance and its impact in estimating power and delay of CMOS digital circuits with low supply voltage," in *ISLPED*, 2000, pp. 228–230.

[30] M. Simard-Normandin, "Channel length dependence of the body-factor effect in NMOS devices," *IEEE TCAD*, vol. 2, no. 1, pp. 2–4, Jan 1983.

[31] P. Flatresse, B. Giraud, J. Noel, B. Pelloux-Prayer, F. Giner, D. Arora, F. Arnaud, N. Planes, J. Le Coz, O. Thomas, S. Engels, G. Cesana, R. Wilson, and P. Urard, "Ultra-wide body-bias range LDPC decoder in 28nm UTBB FDSOI technology," in *ISSCC*, Feb 2013, pp. 424–425.