Design of Mixed Gates for Leakage Reduction

Frank Sill
College of CSEE
University of Rostock
Germany
frank.sill@uni-rostock.de

Jiaxi You
College of CSEE
University of Rostock
Germany
jiaxi.you@uni-rostock.de

Dirk Timmermann
College of CSEE
University of Rostock
Germany
dirk.timmermann@uni-rostock.de

ABSTRACT

Leakage power dissipation is one of the most critical factors for the overall current dissipation and future designs. However, design techniques for the reduction of leakage power should not decrease design performance. Therefore, an enhanced Dual V_{th} / Dual T_{ox} CMOS approach is presented which applies mixed gates consisting of different transistor types. The paper introduces the new and fundamental idea of different gate types before the various possible configurations are analyzed. This is followed by extraction and exploration of design rules and recommendations. Simulations of modified ISCAS'85 designs show an average leakage reduction of 60 % at constant performance compared to raw designs. This corresponds to an additional reduction of 20 % compared to previous Dual V_{th} / Dual T_{ox} CMOS approaches.

Categories and Subject Descriptors

B.7.2 [**Hardware**]: Integrated Circuits – *design aids*

General Terms

Algorithms, Performance, Design

Keywords

Leakage currents, Threshold voltage, Mixed Gates, Gate leakage

1. INTRODUCTION

Reduction of power dissipation is a primary concern of current research in the field of integrated circuits. However, the user demand for high portability and long operation time, especially of battery-operated devices, is contrary to the demand for high performance. Both demands have been satisfied for decades by aggressive downscaling of technology parameters. Thereby, capacitive load per logic gate could be reduced which has allowed ever higher performance as well as decreased dynamic power consumption. Unfortunately, influence of short channel and tunneling effects has increased exponentially which has resulted in dramatically increased leakage currents. It is predicted that power dissipation due to leakage currents will be up to 50 % of the overall power consumption [1]. This development paired with the increasing logic complexity of integrated circuits also intensifies power density issues.

Considering power savings during idle mode is a very common approach to reduce leakage currents. An example of such an approach is the deployment of so-called sleep transistors to disconnect supply voltage from idle modules [2]. Another attempt, called Minimum Leakage Vector, employs special input vectors to the gate inputs [3]. This technique exploits the impact of various input vectors on gate leakage. Yuan et. al. improved the idea by inserting additional gates [4]. As the supply voltage has significant impact on delay and power consumption, Maken et. al. proposed to scale supply voltage according to required performance [5]. Another widely used approach is to apply two device types that vary in their threshold voltage V_{th} [6] or gate oxide thickness T_{ox} [7]. Thus, devices differ in performance and leakage currents. These so-called Dual V_{th} CMOS (DVTCMOS) and Dual T_{ox} CMOS (DTOCMOS) design techniques apply fast gates in critical paths and slower gates with lower leakage are used in non-critical paths. Unlike these two approaches on gate level, Wei et. al. [8] present the DVTCMOS approach based on transistor level. That is, single transistors in the design are replaced by transistors with high V_{th} or low V_{th}, respectively.

GLSVLSI'07, March 11–13, 2007, Stresa-Lago Maggiore, Italy.

The general problems of common leakage reduction techniques are the need for additional devices (e.g. sleep transistors) and the reduction of only one component of leakage. Moreover, transistor level approaches are not applicable for standard cell designs and require long calculation time. Further, gate level DVT-/DTOCMOS methods do not offer the best possible solution as the number of gate types limits the improvement. This paper presents an enhanced DVT-/DTOCMOS approach and analyzes the configuration of applied gate types. After introducing the basics in section 2, section 3 presents the *Mixed Gates* approach. Section 4 concentrates on the analysis, and section 5 gives simulation results.

2. PRELIMINARIES

In the following, a brief description is given to understand the basic idea of DVT-/DTOCMOS approaches and the correlations of transistor parameters in nanometer technologies with gate length below 100 nm.

2.1 Delay

The delay t_d of CMOS devices can be approximated as:

$$t_d \propto \frac{C_{load} V_{DD}}{\mu\left(\varepsilon_{ox} / T_{ox}\right) \cdot \left(W_{eff} / L_{eff}\right)\left(V_{DD} - V_{th}\right)^{\alpha}} \qquad (1)$$

C_{load} labels the load capacitance of the gate, V_{DD} is the supply voltage, μ and ε_{ox} correspond to physical constants (electron surface mobility and gate dielectric constants of gate oxide, respectively), W_{eff} and L_{eff} are the effective gate width and length, T_{ox} labels the thickness of the gate oxide layer, V_{th} is the threshold voltage, and α is the velocity saturation index [9]. Further, V_{th} can be modeled as [10]:

$$V_{th} = V_{th0} + \gamma' \cdot \sqrt{NDEP} \cdot T_{ox} V_{bs} - \eta' \frac{T_{ox}}{L_{eff}^{2} \cdot \sqrt{NDEP}} V_{ds} \qquad (2)$$

V_{th0} is the zero-bias threshold voltage, V_{bs} is the bulk-source voltage, V_{ds} is the drain-source voltage, γ' is the body-bias coefficient, η' is the drain induced barrier lowering (DIBL) coefficient, and *NDEP* labels the channel doping concentration.

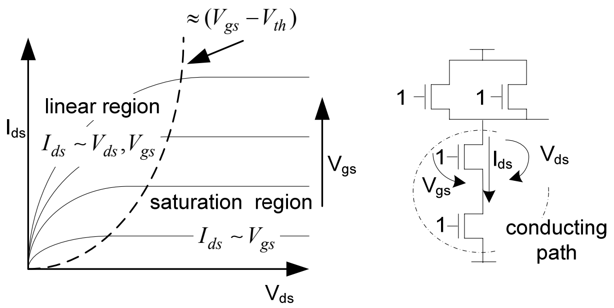


Figure 1. Drain-to-source current (of a NMOS transistor) in depen-dency of gate-source Vgs and drain-source Vds voltages, and a conducting path inside a NAND2 gate

The delay of a gate results from the drain-to-source current of the transistors inside the conducting path (see figure 1). Here, conducting path means the path inside the gate from the output to V_{DD} (for a rising output signal) or from the output to GND (for a falling output signal). The drain-to-source current I_{ds} depends on the gate-source V_{gs} and the drain-source V_{ds} voltages, whereas the strength of the drain-to-source current can be grouped into three regions:

- the *sub-threshold region* ($V_{gs} < V_{th}$), in which I_{ds} is equal to the sub-threshold current I_{sub} (which will be explained in the next subchapter)
- the *linear region* ($V_{ds} < P_v[V_{gs} - V_{th}]^{\alpha/2}$), in which I_{ds} depends on V_{gs} and V_{ds}
- the *saturation region* ($V_{ds} > P_v[V_{gs} - V_{th}]^{\alpha/2}$), in which I_{ds} depends on V_{gs} only

Here, P_v is a technology parameter [9]. Figure 1 depicts the drain-to-source current in the linear and saturation region.

2.2 Leakage

Ideally, CMOS gates draw no current or rather dissipate no power when idle. Unfortunately, this is not true for real gates. The strongest impact originates from sub-threshold current I_{sub} and gate oxide current I_{gate}. The former is the current between source and drain of a transistor when the device should in fact be cut off ($V_{gs} < V_{th}$). A commonly used approximation of sub-threshold current I_{sub} is [10]:

$$I_{sub} = I_0 \cdot \frac{\sqrt{NDEP}}{L_{eff}} \cdot e^{\left(\frac{q}{nkT}\left(V_{gs} - V_{th}\right)\right)} \cdot \left(1 - e^{\left(-\frac{kT}{q} V_{ds}\right)}\right) \tag{4}$$

Here, I_0 is the zero-bias current, T is the operating temperature, n is the sub-threshold swing coefficient, V_{gs} is the gate-source voltage, k is Boltzmann's constant, and q corresponds to the charge of an electron. Following from this, I_{sub} highly depends on V_{th}, which is most sensitive to technology parameters T_{ox} and *NDEP*.

The second important contributor of leakage is the gate oxide current I_{gate} which bases primarily on tunneling currents through the gate's dielectric. Physically, these are the direct tunneling (DT) current in the gate-channel region and the edge direct tunneling current (EDT) in gate-drain/gate-source overlapping regions. Current density for both direct tunneling currents can be described as [12]:

$$J_{DT} = A \cdot \left(\frac{V_{ox}}{T_{ox}}\right)^2 \cdot exp\left(-B\left[1 - \left(1 - \frac{V_{ox}}{\phi_{ox}}\right)^{3/2}\right] \cdot \frac{T_{ox}}{V_{ox}}\right) \tag{5}$$

V_{ox} is the potential drop across the gate oxide, Φ_{ox} is the barrier height for the tunneling particle (electron or hole), and A and B are physical parameters. In contrast to sub-threshold current, gate oxide current is most sensitive to T_{ox} only.

2.3 Dual Vth / Dual Tox CMOS

Data signals traverse integrated circuits through different paths of logic gates whereas start- and endpoints of these paths are marked by sequential elements like registers. The maximum frequency to clock the registers is determined by the path with the longest propagation delay, called critical path. Thus, it is possible to trade off delay for leakage in all other, non-critical paths. Such an attempt is exploited by Dual V_{th} CMOS (DVTCMOS) und Dual T_{ox} CMOS (DTOCMOS) techniques. Therefore, both approaches offer various gates for the same logical function that differ in evaluation delay and leakage.

As shown in the previous section and in [11] the transistor's threshold voltage V_{th} and thickness of the gate oxide layer T_{ox} mainly determine delay and leakage. So, DVT- and DTOCMOS approaches apply fast gates with transistors that have low V_{th} (LVT gates) or low T_{ox} (LTO gates), respectively. Further, gates with same logical functions but consisting of transistors with high V_{th} (HVT gates) or high T_{ox} (HTO gates) are employed. These gates offer slower evaluation but also decreased leakage currents. Finally, to reduce the design's overall leakage currents at constant performance, as many HVT or HTO gates as possible are applied to the non-critical paths so that the delay of the paths does not exceed the critical path delay. The critical path consists solely of LVT or LTO gates [6][7]. In the following, the combination of DVTCMOS and DTOCMOS is called DVTO. Hence, DVTO approaches apply transistors with different T_{ox} and V_{th}, whereas the gates are classified as LVTO and HVTO gates.

3. MIXED GATES

This section describes the fundamental idea of *Mixed Gates* and how the different transistor and gate types were derived. Additionally, a novel scheme to allocate the different gate types in a design is presented.

3.1 Fundamental Idea

The idea of the *Mixed Gates* design technique is the combination of advantages from gate level and transistor level approaches. That is, high accuracy and applicability for a standard cell design flow. In addition, the *Mixed Gates* approach merges both DVTCMOS and DTOCMOS approaches.

As shown in section 2 and in [11], transistor leakage as well as delay highly depends on V_{th}, T_{ox}, and *NDEP*. Against this background, our previous work studied new transistor types. Transistor models of a predictive 65 nm technology were used as base technology, and two types of transistors were defined by extensive simulations [13][14]. The first type, fast switching transistors *L-Vt/To* with high leakage currents, uses low T_{ox} and low *NDEP*, which results in low V_{th} and short delay. The second transistor type *H-Vt/To* has high T_{ox} and high *NDEP*, which results in high V_{th}. The latter transistors switch slower but have lower gate oxide leakage as well as reduced sub-threshold leakage. HVTO gates consisting solely of *H-Vt/To* transistors and LVTO gates consisting solely of *L-Vt/To* transistors have been imple-

Table 1. Classification of gate types that are apllied by the Mixed Gate design technique

	HVTO	MG	F-MG
Delay	High	Medium	Low
Leakage	Low	Medium	High
Applied transistor types	H-Vt/To	H-Vt/To and few L-Vt/To	L-Vt/To and few H-Vt/To

mented (see figure 2). At this stage, leakage reduction based on the idea of DVTO approaches can be performed.
However, the *Mixed Gates* approach offers two additional gate types, which are called standard Mixed Gates (MG) and Fast Mixed Gates (F-MG). F-MG gates contain few *H-Vt/To* transistors at adequate positions so that delay is equal to corresponding LVTO gates but leakage is lower. Therefore, LVTO gates are not used any further because they can be replaced by F-MG gates without any drawback on performance due to equal delay but lower leakage. Finally, MG gates contain mainly *H-Vt/To* transistors, so that gate delay is in between the delay of F-MG and HVTO gates. Figure 2 depicts four variations of a NOR2 gate and table 1 classifies mixed gate types qualitatively.

When applying the *Mixed Gates* approach, a design is initially implemented with F-MG gates only. Then, as many gates as possible are exchanged by appropriate MG and HVTO gates so that maximum delay is not increased. That is, critical paths consist solely of F-MG gates whereas all other paths apply a combination of the three gate types so that path delays do not exceed critical path delay. Thus, the *Mixed Gates* approach offers three advantages:

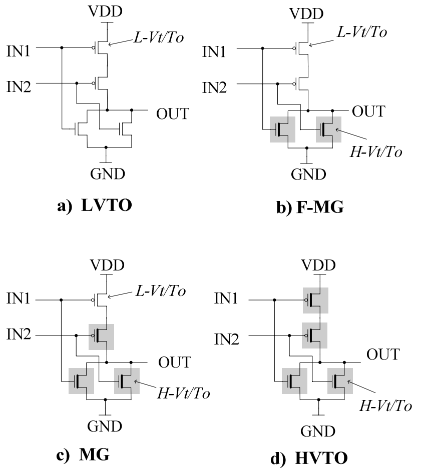
- larger amount of low leakage devices within a design than possible in DVTO approaches
- reduction of both main sources of leakage currents: gate oxide leakage I_{gate} and sub-threshold leakage I_{sub}
- no drawback on performance compared to the original designs which consist of fast and high leakage devices

3.2 Gate Type Allocation Algorithm

An allocation algorithm is needed to apply different gate types at appropriate positions so that leakage is reduced without affecting system delay. The developed algorithm bases on weighting each gate of the design. The weight factor of each gate is calculated from its leakage, delay, and slack. The latter is the time by which a gate can be slowed down without worsening design performance. In the initialization phase, each gate is set to the F-MG type. Based on the weight, MG gates or a HVTO gates are applied if possible. Subsequently, pin reordering follows to consider dependency of gate leakage on input values. The interested reader is referred to [7] for a detailed description of the pin reordering and the applied algorithm [14].

4. ANALYSIS OF DIFFERENT GATE CONFIGURATIONS

The main problem of the *Mixed Gates* approach is the realization of a gate library with F-MG and MG gate types. As mentioned in section 3.1, F-MG gate type must have the same delay as the LVTO gate version while MG gate type requires the best trade-off between delay and leakage. Additionally, the input capacities of all gate types should be nearly the same to prevent increase of dynamic power dissipation. Unfortunately, implementation and comparison of all possible combinations of different transistor types within a gate is quite costly. For example, a gate consisting of six transistors offers 26 = 64 possible configurations. Hence, the extraction of configuration recommendations for both gate types is necessary. In the following, rules and recommendations will be extracted, starting from the implementation and analysis of different gate type realizations.



a) LVTO **b) F-MG**

c) MG **d) HVTO**

Figure 2. Different implementations of a NOR2 gate: LVTO, HVTO, and *Mixed Gates*

4.1 Test Environment

The test environment for analyzing different gate configurations models a typical case. Thus, each input signal, which is connected to the tested gate, has an additional load of three LVTO inverters. These inverters have about the same input capacity as the inputs of the tested gate. Further, the gate's output has a load of four LVTO inverters (FO4). The new gates are derived from a standard LVTO gate library with each gate sized for equal maximum rising and falling slopes. The used transistors are based on a modified 65 nm technology [13][14]. For each gate, different combinations of *L-Vt/To* and *H-Vt/To* transistors were analyzed. This includes all possible realizations of mixed transistor stacks. For each test case, delay and leakage of all feasible input combinations were simulated. Starting from this, average leakage and maximum delay of the tested gate configuration could be calculated.

4.2 Design Rules for Mixed Stacks

Within the observed CMOS gates two kinds of arranged transistor structures exist: parallel and stacked. It is recommended to use same transistor types in parallel structures. This follows from the fact that the maximum evaluation delay t_{max} of such parallel structures results from only one conducting transistor independent of the other parts. Thus, parallel structures offer only two t_{max} at constant input capacity: a high t_{max}, if at least one transistor is *H-Vt/To* type, and a low t_{max}, if all transistors are *L-Vt/To* type.
In contrast, transistors stacks offer some more possibilities. Hence, following from the results of all simulated test cases two design rules for transistor stacks have been extracted:

Delay rule: *Within mixed stacks, the LL-Vt/To transistors have to be placed as close closeas possible to the gate output to achieve best results for the delay.*

Leakage rule: *Within mixed stacks, the H-Vt/To transistors have to be placed at the end of stack (away from the output) to achieve best leakage results.*

Delay rule

In all simulated stacks the maximum delay occurs if all stacked transistors switch at the same time. This could be changed only, as input and output loads were reduced to one inverter, respectively. Further, the input slope delay was set to one third of the LVTO inverter delay which has the minimum delay within the library. In this theoretical case, the maximum delay occurs if only the transistor farthest from the gate output is switching. This is based on the internal capacities within the stack which have to be (dis-)charged additionally to the output load. For the further analysis this case was ignored, because it is very specific, and happens with a very low probability only. Thus, the estimation of the *delay rule* is based on complete switching stacks (figure 3).

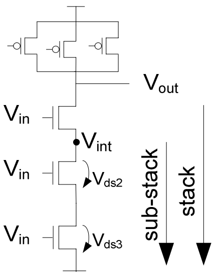


Figure 3. Stack and sub-stack of a NAND3

Before switching, the internal nodes within the stack (eg. V_{int} in figure 3) have a potential close to zero. Thus, only the topmost transistor has a high drain-source voltage V_{ds}. Hence, the topmost transistor works in saturation region only ($V_{ds} > P_v[V_{gs} - V_{th}]\alpha/2$, see section 2.1) when all input voltages V_{in} of the stack start to change. However, the other transistors in the sub-stack (see figure 3) work in linear region ($V_{ds} < P_v[V_{gs} - V_{th}]\alpha/2$, see section 2.1) [15]. The conductance g of a transistor in linear mode is proportional to:

$$g \sim (V_{gs} - V_{th} - V_{ds}) \qquad (4)$$

Further, the current through the stack I_{stack} is proportional to gate output voltage V_{out} and stack conductance g_{stack}:

$$I_{stack} \sim V_{out} \cdot g_{stack} = V_{out} \cdot \left(\frac{1}{g_{top}} + \sum_{sub\_stack} \frac{1}{g_{sub\_stack\_i}} \right)^{-1} \qquad (5)$$

g_{top} denotes conductance of the topmost transistor of the stack. It can be observed that a higher conductance leads to a higher I_{stack}. Further, a higher current through the stack results in faster discharge of the gate output node and, consequently, in lower gate delay. Thus, an increasing conductance of the topmost transistor or of the sub-stack leads to a shorter delay.

Simulations indicate the best configuration of a mixed three-transistor stack is "LLH", which consists of two upper *L-Vt/To* transistors and one bottom *H-Vt/To* transistor. The behavior has to be compared with the two other configurations which applies one topmost *H-Vt/To* transistor ("HLL") and one *H-Vt/To* transistor in the middle of the stack ("LHL"), respectively.

LLH vs. HLL

The internal node voltage V_{int_LLH} in "LLH" configuration is always higher than V_{int_HLL} in "HLL" configuration. This is due to better conductance of *L-Vt/To* transistors compared to *H-Vt/To* transistors. Thus, the drain-source voltages V_{ds2} and V_{ds3} within the sub-stack in "LLH" configuration are higher as in "HLL" configuration. This leads to lower conductance of the sub-stack in version "LLH" (see equation 5 and figure 4). This effect is intensified by the *H-Vt/To* transistor within the sub-stack. However, conductance of the whole stack in "LLH" configuration is higher than "HLL" (see figure 4) due to high conductance of the topmost transistor in version "LLH". This effect could also be verified at bigger stacks.

LLH vs. LHL

Internal node voltages V_{int} of "LLH" and "LHL" configurations are nearly the same. However, the conductances g_{sub_stack} of the sub-stacks differ. The bottommost transistor of a stack has highest V_{gs} within the stack, because its source is directly connected to GND. Hence, compared to the *H-Vt/To* transistor in "LHL" the conductance of the *H-Vt/To* transistor in "LLH" is higher (see equation 5). This effect has higher influence than better conductance of the *L-Vt/To* transistor in the sub-stack of "LHL" compared to the bottommost L-Vt/To transistor of "LLH" version. Hence, compared to "LHL" configuration g_{sub_stack} and consequently g_{stack} of the "LLH" version are higher (see figure 4).

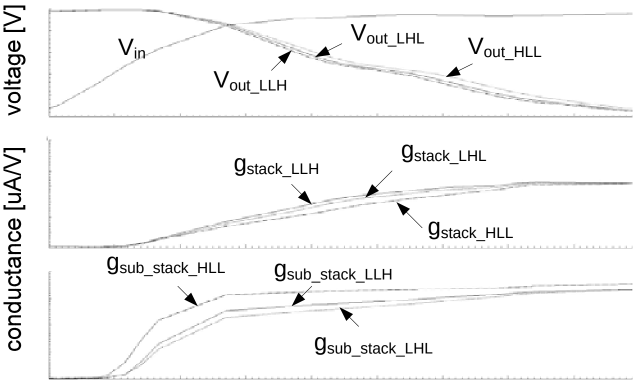


Figure 4. Comparison of output voltage Vout**, conductance of the whole stack** gstack**, and conductance of the sub-stack below the top transistor** gsub_stack **of different configurations of 3-NMOS stacks in a NAND3. The letters indicate the positions (top-to-bottom) of** *H-Vt/To* **("H") and both** *L-Vt/To* **transistors ("L") within the stack.**

Leakage rule

As expected, simulations indicate that average sub-threshold current I_{sub} is nearly the same in stacks with same amount of transistor types. This bases on the fact, that in a stack the drain-source current through all transistors is the same (Kirchhoff's law). As I_{sub} occurs only between drain and source, it makes almost no differences where the different transistor types are placed.

In contrast, the gate oxide current I_{gate} strongly depends on transistor positions inside the stacks. As described in section 2.2, gate tunneling current occurs if voltage difference exists between gate and source or gate and drain (edge direct tunneling, EDT) or gate and bulk (direct tunneling, DT). Table 2 depicts the gate leakage current of all possible combination of the terminal's potential of a single *L-Vt/To* NMOS transistor, whereas the bulk potential Φ_{bulk} is always at 0 V. It can be observed, the highest gate oxide current I_{gate_max} occurs when the gate potential Φ_{gate} is V_{DD} and the potential of source and drain Φsource,drain = 0 V. In this case, there is a gate oxide current between gate and drain, gate and source, and gate and bulk. I_{gate} decreases to 2/3 and 1/3 of I_{gate_max} when Φ_{gate} = 0 V, and Φsource and Φdrain are at V_{DD}. This decrease bases on the sub-threshold current, which occurs when Φ_{gate} is lower than the threshold voltage (see section 2.2). Thus, there are two paths from drain or source to ground. When the gate terminal, and the drain or the source terminals have a potential of V_{DD} the gate is conducting (see section 2.1). It can be observed that even in this case a gate oxide current occurs. However, when all potentials are near V_{DD} only gate-bulk current occurs, which is very low. Hence, each transistor within a stack has different average I_{gate} depending on its terminal potentials.

Table 2. Absolute I_{gate} of an NMOS *L-Vt/To* transistor in dependency of terminal potentials (Φ_{bulk} = 0 V)

Φ_{source}	Φ_{drain}	Φ_{gate}	$\|I_{gate}\|$
0	0	0	0 nA
0	0	V_{DD}	24.8 nA
0	V_{DD}	V_{DD}	15.04 nA
V_{DD}	V_{DD}	0	15.04 nA
0	V_{DD}	0	7.83 nA
V_{DD}	0	0	7.83 nA
V_{DD}	V_{DD}	0	15.66 nA
V_{DD}	V_{DD}	V_{DD}	1.1e-9 nA

Table 3 compares the most dominating gate-oxide leakage currents inside a stack of three NMOS transistors for all input vectors. From this follows, the highest gate oxide current happens when all inputs are at V_{DD} potential. Further, it can be observed that highest average gate oxide current occurs at the bottom transistor (T3). This is based on the fact that the source terminal is always at 0 V. Thus, there occurs always the highest gate current when the gate potential of T3 is V_{DD}. In contrast, the source or drain potential of the both other transistors is not always at 0 V when its gate potential is V_{DD}. Hence, bottom transistor should be of *H-Vt/To* type to achieve best improvements.

4.3 Mixed Gates Design Recommendation

Following from our simulations and design rules, recommendations for the design of *Mixed Gates* can be formulated:

MG gates

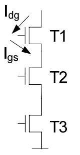
Parallel transistors within MG gates should consist of *H-Vt/To* transistors only, while stacks should be realized as mixed stacks using the rules given above. Before sizing, the maximum delay t_{max} of parallel structures is longer compared to stacked structures. Therefore, transistor widths are adapted during sizing for balanced maximum falling and rising slopes whereas total input capacity remains constant. Further, amount of *L-Vt/To* transistors within stacks should be varied until desired t_{max} is reached.

F-MG gates

At PMOS transistors gate leakage I_{gate} is one decade smaller than corresponding sub-threshold leakage I_{sub} [7]. This justifies formulation of two recommendations for implementing F-MG gates: all parallel transistors are *H-Vt/To* type and stacked transistors are *L-Vt/To* type or, alternatively, stack includes less *H-Vt/To* transistors and parallel transistors are *L-Vt/To* type. The advantage of the first solution is strongly reduced I_{sub} when all parallel transistors are non-conducting. Unfortunately, this case occurs only at one input vector. Else, I_{sub} is the same as in corresponding LVTO gates. However, I_{gate} can be reduced in every case at each *H-Vt/To* transistor. The second solution leads to higher amount of cases in which I_{sub} is reduced while reduction of I_{gate} decreases. Hence, the first solution is recommended if I_{gate} of stack is comparable to I_{sub}, like in NAND gates. The second solution should be applied if I_{gate} of the stacks can be ignored, like in NOR gates.

Table 3. Classification of most dominationg gate oxide currents I_{gate} of a NMOS stack for all input vectors (↓ low, → medium, ↑ high I_{gate})

Vector	T1	T2	T3
000	I_{gate_dg}: ↓	-	-
001	I_{gate_dg}: ↓	-	I_{gate_gs}: ↑
010	I_{gate_dg}: ↓	$I_{gate\_gs} + I_{gate\_dg}$: →	-
011	I_{gate_dg}: ↓	I_{gate_gs}: ↑	I_{gate_gs}: ↑
100	-	I_{gate_dg}: ↓	-
101	-	I_{gate_dg}: ↓	I_{gate_gs}: ↑
110	-	-	I_{gate_dg}: ↓
111	I_{gate_gs}: ↑	I_{gate_gs}: ↑	I_{gate_gs}: ↑



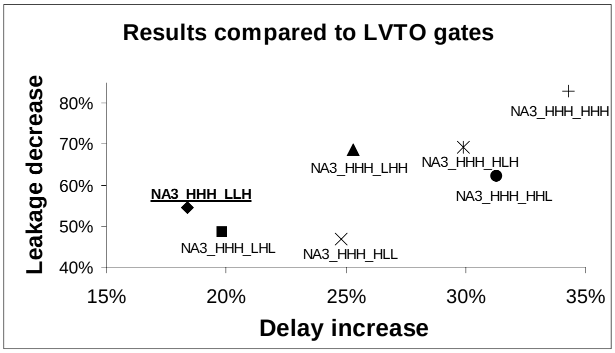


Figure 5. Comparison of possible MG realizations for a NAND3 (NA3) gate. The underlined configuration was chosen for the library. The first three letters indicate the type of the PMOS transistors and the last three the type of the NMOS transistors. There, 'H' means *H-Vt/To* type, while 'L' means *L-Vt/To*. The ordering starts with highest devices in the stacks.

Unfortunately, it is impossible to realize F-MG gates with identical input capacities compared to LVTO gates, if t_{max} remains constant. Thus, a marginal input capacity penalty has to be accepted. However, simulations indicated only minor impact on load and dynamic power dissipation.

4.4 Library Creation

Based on our rules and recommendations a gate library of ten standard gates was created. Cadence Virtuoso Circuit Optimizer was used to achieve the best sizing for each gate. Table 4 shows the options for the Circuit Optimizer of each group. To verify the benefit of our rules and recommendations all other gate configurations were also implemented with same optimizer options (e.g., see figure 5 and figure 6). MG

Table 4. Options of each group for the circuit optimizer

LVTO	- Each input has the same C_{in} - maximum delay of 100 ps for NAND / NOR - maximum delay of 150 ps for AND / OR - balanced maximum delays of falling and rising slopes (± 10 %)
HVTO	- same C_{in} of each input (including internal loads) as corresponding LVTO type - balanced maximum delays of falling and rising slopes (± 10 %)
MG	- same C_{in} of each input (including internal loads) as corresponding LVTO type - balanced maximum delays of falling and rising slopes (± 10 %)
F-MG	- nearly same C_{in} of each input (including internal loads) as corresponding LVTO type (± 5 %) - same maximum delay as corresponding LVTO type - balanced maximum delays of falling and rising slopes (± 10 %)

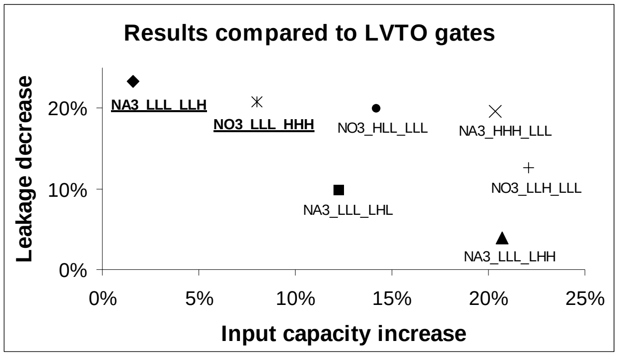


Figure 6. Different implementations of F-MG for a 3-input NAND (NA3) and NOR (NO3) gate. The underlined configuration was chosen for the library.

and F-MG analysis results indicate the possibility to realize mixed gates with desired requirements. The input load of the chosen F-MG gates increased by less then 5 % while delay remained constant. Leakage decreased by 20 %. MG gates are viable with medium delay compared to corresponding HVTO and LVTO gates while leakage decreased by over 50 %. Hence, the extracted rules and recommendations could be approved.

5. RESULTS

To verify the *Mixed Gates* approach, ISCAS'85 benchmark designs were implemented based on our gate library [16]. To cope with the lack of an additional sizing algorithm the circuits were synthesized with limited load. At first, all designs were designed with LVTO gates (LVTO version). The second version (DVTO version) contains LVTO gates in critical paths and HVTO gates in non-critical paths. This implementation corresponds to mixed standard DVT-/DTOCMOS design techniques. The third implemented type of each ISCAS'85 design is the *Mixed Gates* version with F-MG, MG, and HVTO gates.

Each design version was simulated with Synopsys HSpice to determine leakage and power dissipation for an active mode.

Circuit leakage was measured for a signal probability of 0.5 for every design input. In active mode, designs worked with a frequency of *1 GHz* and an activity *α=25%.* Results (see figure 7) indicate that leakage of the *Mixed Gates* versions is in average 60 % lower than corresponding LVTO versions and roughly 20 % lower than DVTO versions. The dynamic power dissipation of the Mixed Gates designs is in average 10 % lower as in corresponding LVTO designs. This is based on reduced effects of glitches and hazards that are caused by unbalanced paths. The dynamic power dissipation of *Mixed Gates* designs is in average 1 % higher compared to DVTO, caused by the slightly higher input capacity of F-MG gates compared to LVTO gates.

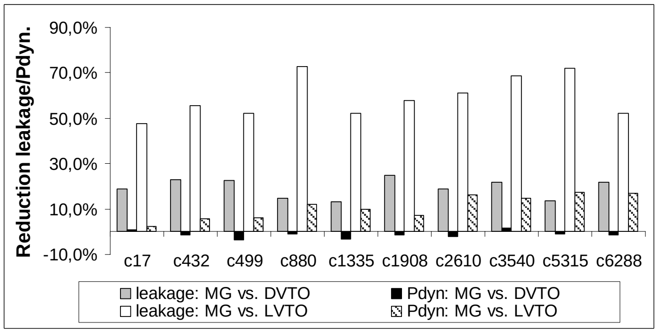


Figure 7. Results of simulated ISCAS'85 designs

6. CONCLUSIONS

This paper proposes the *Mixed Gates* design technique for leakage reduction at constant performance. Our approach applies gates consisting of different transistor types. This results in gates with same logical function but different gate delay and leakage power dissipation. Different gate configurations were analyzed and design rules and recommendations were extracted. The comparison between raw LVTO and *Mixed Gates* designs shows an average leakage reduction of 60 % and an average dynamic power reduction of 10 %. The leakage could be reduced by 20 % in average compared to combined standard Dual V_{th} CMOS and Dual T_{ox} CMOS design methods. The dynamic power dissipation increased only slightly and, thus, does not counteract the achieved power reductions.

7. REFERENCES

[1] Kim, N.S., et. al.: Leakage Current: Moore's Law Meets Static Power, *IEEE Computer*, p. 68, no. 12, 2003.

[2] Anis, M., and Elmasry, M. in *Multi-Threshold CMOS Digital Circuits*, Kluwer Academic Publishers (2003).

[3] Tsai, Y., et. al.: Characterization & modeling of run-time techniques for leakage reduction. *Tr. VLSI Syst. vol. 12*, 2004.

[4] Yuan, L. and Qu, G.: Enhanced leakage reduction Technique by gate replacement. *42nd DAC*, San Diego, 2005.

[5] Maken, P.: et. al.: A Voltage Reduction Technique for Digital Systems, *ISSCC*, 1990.

[6] Sundararajan, V. and Parhi, K.: Low Power Synthesis of Dual Vth CMOS VLSI Circuits, *ISPLPED*, 1999.

[7] Sultania, A.K., Sylvester, D., Sapatnekar, S.: Transistor and Pin Reordering for Gate Oxide Leakage Reduction in Dual T_{ox} Circuits, *22nd ICCD,* San Jose, USA, 2004.

[8] Wei, L., et. al. : Mixed-V_{th} (MVT) CMOS Circuit Design Methodology for Low Power Applications, *36thDAC,* 1999.

[9] Sakurai, T. and Newton, R.: Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas, *IEEE JSSC*, 25 (2), Apr. 1990.

[10] Hu, S., et. al.: *Berkeley short channel IGFET model,* Dpt. of EECS, University of California, Berkeley, 2005.

[11] Mukhopadhyay, S. and Roy, K.: Modelling and Estimation of Total Leakage Current in Nanoscaled CMOS Devices Considering the Effect of Parameter Variation, *ISLPED'03*, Seoul, Korea, 2003.

[12] Schuegraf, K. and Hu C. Hole injection SiO break down model for very low voltage life time extrapolation, *IEEE Trans. Electron. Devices*, vol.41, pp. 761–767, 1994.

[13] Cao, Y., et. al.: New paradigm of predictive MOSFET and interconnect modeling for early circuit design, *CICC*, 2000.

[14] No reference due to blind review

[15] Bisdounis, L., et. al.: Modeling the dynamic behavior of series-connected MOSFETs for delay analysis of multiple-input CMOS gates, ISCAS, USA, 1998.

[16] Hansen, M., Yalcin, H., Hayes, J. P. Unveiling the ISCAS-85 Benchmarks: A Case Study in Reverse Engineering, *IEEE D&T*, vol. 16, no. 3, pp. 72-80, July-Sept. 1999.

[17] Lee, D., Kwong, W., Blaauw, D., and Sylvester, D. Analysis and minimization techniques for total leakage considering gate oxide leakage, *40th DAC*, Anaheim, USA, 2003