Automatic Device Design Optimization with TCAD Frameworks

M. Stockinger and S. Selberherr

Institute for Microelectronics, TU Vienna
Gusshausstrasse 27–29, A-1040 Vienna, Austria, selberherr@tuwien.ac.at

ABSTRACT

A design optimization method is presented which utilizes automatic optimization capabilities within TCAD frameworks. This method is applied to doping profile optimizations of ultra-low-power CMOS transistors with 0.25 and 0.1 μm gate lengths. Two different performance goals are utilized, to maximize the drive current of an NMOS transistor and to minimize the gate delay time of a CMOS inverter stage. These optimizations result in an asymmetric doping profile with a channel peak near the source. Gaussian functions are used to simplify the doping structure without much of a performance loss. The inverter speed of the 0.1 μm technology is improved by almost 100% compared to an inverter with uniformly doped devices delivering the same off-state leakage current.

Keywords: Optimization, CMOS, TCAD, Simulation, Ultra-Low-Power.

1 INTRODUCTION

During the last years the use of TCAD tools for process and device design/analysis is becoming the backbone of the cost efficient and high-performance production of microchips. Due to the constantly increasing computational power of computer systems the simulation times are becoming smaller and smaller enabling the use of TCAD tools on a very large scale.

Especially for complex optimization tasks they have gained much attraction because manually controlled simulations are not suitable. TCAD frameworks have been developed which offer automatic optimizations and require no user interaction during the actual optimization process.

Such frameworks are of special significance for device design purposes where it is necessary to improve a certain performance metric by variation of given design parameters. Amongst all possible optimization strategies, an iterative method where the performance metric is gradually improved using gradient information of the design parameters, has delivered excellent results [1].

In this work the drive and inverter performance of ultra-low-power CMOS transistors is improved by doping profile optimizations performed on two different device generations. A very general approach is taken to define the doping profiles which results in a large number of design parameters. This complex optimization task is performed within the TCAD framework SIESTA which is perfectly suited for this purpose [2].

Though our very general approach ignores manufacturability issues in the first place, the optimization results will give valuable ideas to improve existing technologies or might even encourage the development of completely new device concepts.

2 DESIGN SETUP

Two different device generations are considered in this work, as listed in Table 1. The supply voltages match with the requirements for ultra-low-power applications.

Table 1: Key parameters of the two device generations considered in this work

Generation	L_g	T_{ox}	V_{dd}
A	0.25 μm	5.0 nm	1.5 V
B	0.10 μm	2.5 nm	0.9 V

A quite simple device architecture is used. It is a planar structure with SiO_2 source and drain spacers and an SiO_2 gate insulator (Fig. 1).

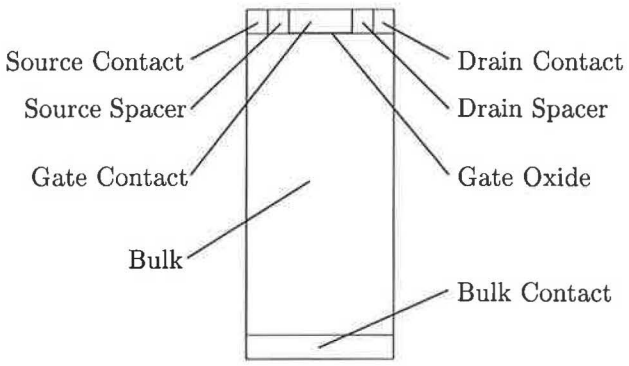

Figure 1: The used device architecture

The source and drain doping profiles are modeled using Gaussian functions with a maximum surface concentration of 10^{20} cm^{-3}. The exact values of the junction depths depend on the substrate background doping. Rough values are 50–120 nm for the 0.25 μm device and 20–48 nm for the 0.1 μm device.

The design space is defined by a discretization of the two-dimensional bulk doping profile of the transistors. The device geometries, supply voltage, and source and drain doping profiles are kept fixed during optimization.

A two-dimensional discretization method is utilized to provide a very general characterization of the doping profiles in the bulk. It covers the area between and under the source and drain wells including the channel, and has the shape of an inverted-T, as shown in Fig. 2. The doping values at each of the grid points are the optimization parameters determining the design space. For the rest of the bulk a constant doping of 10^{15} cm^{-3} is used.

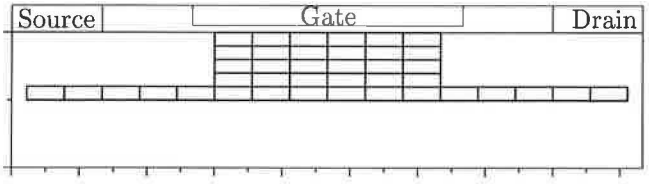

Figure 2: The inverted-T doping discretization

To obtain a smooth two-dimensional doping profile, an interpolation method is utilized between the grid points. This method uses two-dimensional raised-cosine shaped fragments, one placed at each of the grid points, which are superposed in the logarithmic domain. Each of the raised-cosines reaches exactly to the next closest grid points influencing only the area within its neighbors.

3 OPTIMIZATION

The optimization procedure can be illustrated by a black box evaluator which offers a set of optimization parameters at its input and a performance metric, also called "optimization target", and constraints at the output. The goal is to find a set of optimization parameters which deliver the best performance metric while keeping the constraints within a specified range.

For this task the TCAD framework SIESTA is used which utilizes a gradient based optimization method: A nonlinearly constrained optimizer donlp2 supports closed-loop optimizations within this environment [3].

A uniformly doped inverted-T region is used as the initial doping. After the optimization has been setup, it runs automatically without any user interaction required. The procedure is very efficient and stable because SIESTA provides dynamic load balancing and automatic repetition of failed simulation jobs [4].

The optimization loop is shown in Fig. 3. After SIESTA has started the loop with a set of optimization parameters, the device generator reads the parameters, builds the doping profile, and writes the device description. Then device simulations are carried out using MINIMOS-NT [5]. The simulation results are delivered back to SIESTA and used to derive the optimization target and the constraint.

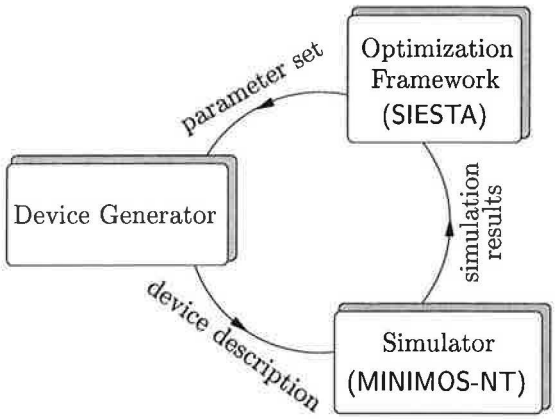

Figure 3: The optimization loop

In this work two different design goals are considered. The first one is to maximize the drive current of an NMOS transistor, the second one is to minimize the gate delay time of a CMOS inverter. While the number of optimization parameters is 64 in case of drive current optimizations, this number is doubled for the gate delay time optimizations because the doping profile of both the NMOS and PMOS devices of the inverter is optimized at the same time. The average leakage current is kept below 1 pA/μm in both cases, which is a realistic value for ultra-low-power applications.

3.1 Drive Current Optimization

The optimization target for drive current optimizations, which will be minimized, is defined as

$$\text{target} = -\frac{I_{\text{on}}}{1\,\mu\text{A}} \quad \rightarrow \text{min.}\,, \tag{1}$$

and the constraint reads

$$\text{constr.} = -\log\left(\frac{I_{\text{off}}}{1\,\text{pA}}\right) \quad > 0 \tag{2}$$

and will be kept above 0 during the optimization. I_{on} denotes the drive current ($V_{\text{g}} = V_{\text{dd}}$) and I_{off} the drain-source leakage current ($V_{\text{g}} = 0$ V) of the transistor.

Usually, a change in channel doping has an exponential-like impact on the leakage current. The use

of a logarithmic transformation reduces this nonlinearity and leads to a better convergence of the optimization procedure.

The evaluation network for drive current optimizations is depicted in Fig. 4. After reading the optimization parameters, the device description of the NMOS is generated containing all necessary data to perform two-dimensional device simulations with MINIMOS-NT for the drive current I_{on} and the leakage current I_{off} in the next steps. The resulting currents are used to evaluate the target and the constraint. This network is embedded into the closed-loop optimization process and its evaluation is initiated by the framework any time the target and constraint values are required for a new set of optimization parameters.

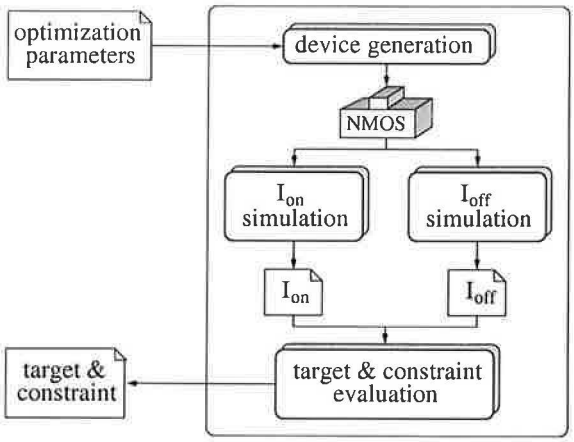

Figure 4: The evaluation network used for drive current optimizations

3.2 Gate Delay Time Optimizations

In order to evaluate the average inverter delay time, an adequate model for one single stage is used (Fig. 5) which bevaves like being part of an infinite inverter chain. It consists of a CMOS inverter and a capacitive load C_L connected to the output which accounts for the gate capacitance of the following stage. Since C_L changes during transition, it is assumed to be voltage dependent. It can be calculated using the input current information of the succeeding stage:

$$C_L(V) = \left. \frac{I_{in}(t)}{dV_{in}(t)/dt} \right|_{V_{in}(t) = V} \quad (3)$$

The interconnect capacitances are neglected, therefore this model represents an ideal case and the resulting delay time will be a lower limit set by the intrinsic quantities of the devices only.

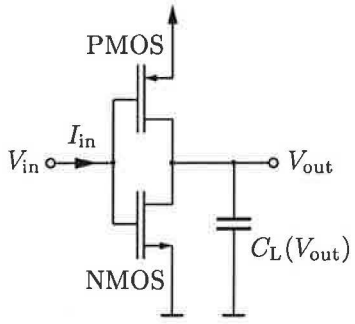

Figure 5: The single stage inverter model

For the actual design goal the optimization target is defined as the average inverter delay time for the on- and off-transitions:

$$\text{target} = \frac{1}{2} \cdot \frac{(t_{d,on} + t_{d,off})}{1\ \text{ps}} \rightarrow \min. \quad (4)$$

The constraint guarantees that the average leakage current stays below 1pA/μm. This time both the NMOS and PMOS leakage currents have to be taken into account:

$$\text{constr.} = -\log\left(\frac{(I_{off,NMOS} + I_{off,PMOS})/2}{1\ \text{pA}}\right) > 0 \quad (5)$$

The evaluation network for gate delay time optimizations is shown in Fig. 6. After reading the doping parameters, the device descriptions of the NMOS and PMOS transistors are produced. Then the inverter model depicted in Fig. 5 is evaluated by mixed-mode transient simulations with MINIMOS-NT for both the on- and off-transitions. Additional input data for the simulator, besides the device descriptions, are the input V-t curves and the C-V curves of the capacitive load C_L which are all taken from a data container. Using the resulting output V-t and input I-t curves of the inverter, the delay times and leakage currents are calculated in a post-processing step. To find the delay times, the time-points when the inverter's input and output V-t curves cross the $V_{dd}/2$ mark are extracted and subtracted from each other. Furthermore, the input V-t and C-V curves for following model evaluations are processed and stored in the data container.

Since the presented optimization procedure features a gradual improvement of the target, the permanent update of the input V-t and load C-V curves provides a self-contained emulation of an infinite inverter chain. At the time when the optimization procedure converges, the input and output curves of the inverter will be self-consistent meaning that the inverter behaves like one stage of a ring oscillator.

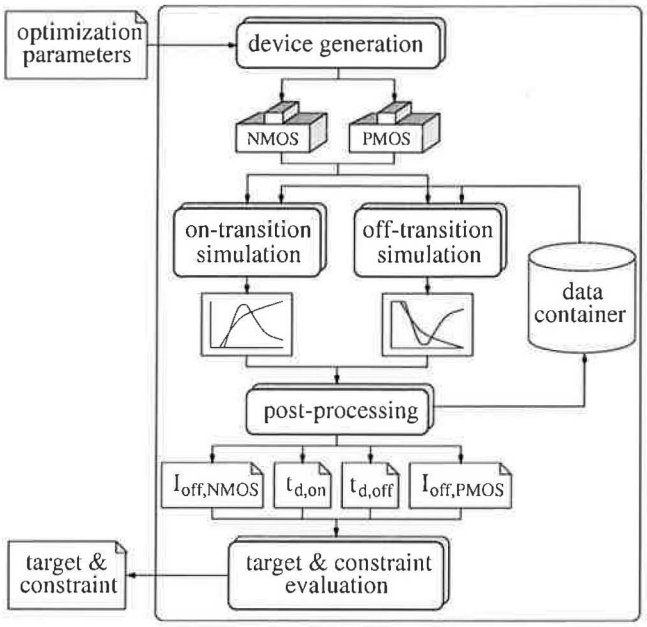

Figure 6: The evaluation network used for gate delay time optimizations

4 RESULTS

To provide an example of the resulting doping profiles, the optimized 0.1 μm NMOS and PMOS profiles for minimum gate delay time are shown in Fig. 7. This quite complex doping structure is examined in a sensitivity analysis to determine how much influence each of the doping parameters has on the performance. The doping regions with very little influence can then be tailored in order to reduce the complexity.

The total relative delay time sensitivity of the NMOS transistor is shown in Fig. 8, the sensitivity of the PMOS transistor looks very similar and is not shown. There is one important region which is located in the channel region, slightly beneath the silicon surface, close to the source well. At this position there are local maxima of the doping concentration as pointed out by the arrows in Fig. 7.

All other optimization results not shown have the same features: The optimum doping profiles contain a doping peak in the channel close to the source and the sensitivity has a clear maximum at this particular place.

It has been shown in [6] that this doping peak sets the threshold voltage of a device and reduces the effective channel length, thereby increasing the drive performance of a transistor. Consequently, the gate delay times are also reduced as the higher drive current can charge and discharge the inverter output node more quickly. An experimental verification of this device structure using a Focused Ion Beam implantation method can be found in [7] where the superior characteristics of a peaking channel doping are pointed out.

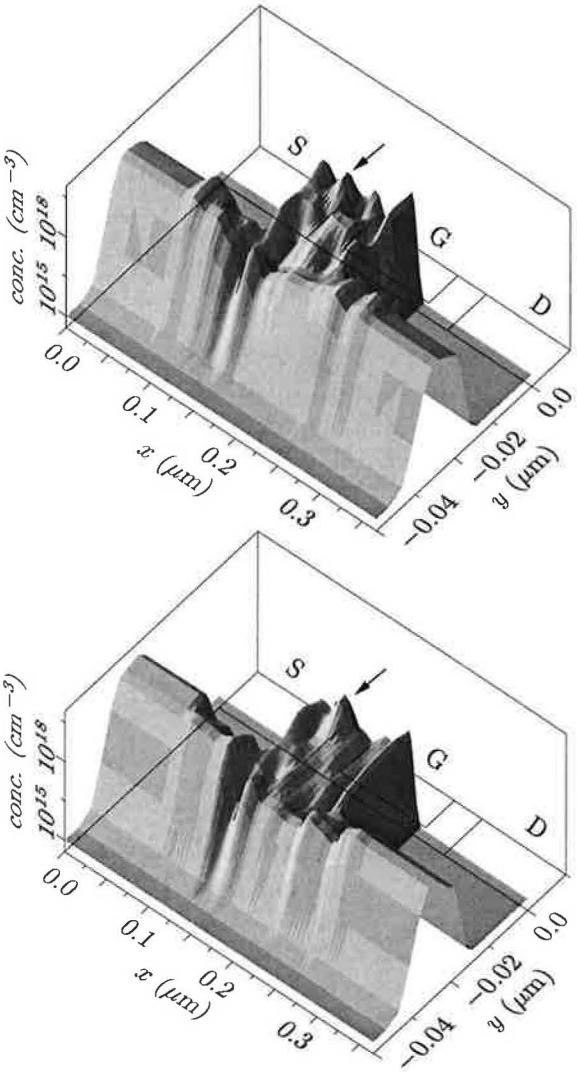

Figure 7: The gate delay time optimization results using the inverted-T structure for the NMOS (top) and PMOS (bottom) transistors with 0.1 μm gate lengths

This very general two-dimensional design approach provides valuable information about how to design the doping profile in order to improve the device performance. It does not assume any a-priori knowledge about a "good" doping profile as not to restrict the optimization possibilities. Anyway, the gained knowledge can be used in a further step to reduce the complexity of the simulation setup by introducing analytical functions to model the doping profiles.

In this work two Gaussian functions are used, one for the channel peak, the other to build a punchthrough-stopper under the source. To provide an initial device, the Gaussian functions are manually fitted to the results obtained from inverted-T structure optimizations. The whole optimization procedure is started again, now with the Gaussian parameters defining the design space.

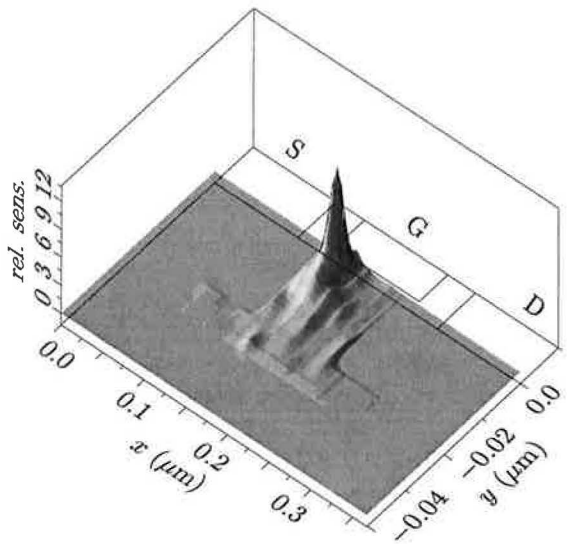

Figure 8: The gate delay time sensitivity of the NMOS transistor with 0.1 μm gate length

The optimized 0.1 μm NMOS and PMOS Gaussian profiles are shown in Fig. 9. The performance gain using this much simpler structure is almost as high as of the fully two-dimensional approach.

5 DISCUSSION

The performance improvements on the basis of drive current and gate delay time optimizations are listed in Table 2 and Table 3, respectively. They refer to uniformly doped devices delivering a leakage current of 1 pA/μm.

Table 2: Performance improvements due to drive current optimizations

	Generation A		Generation B	
doping	I_{on} (μA)	gain	I_{on} (μA)	gain
uniform	258.5	–	130.8	–
two-dim.	373.7	44.6%	224.2	71.4%
Gauss	373.2	44.4%	222.5	70.1%

Table 3: Performance improvements due to gate delay time optimizations

	Generation A		Generation B	
doping	t_d (ps)	gain	t_d (ps)	gain
uniform	53.7	–	72.5	–
two-dim.	34.9	53.9%	36.8	97.0%
Gaussian	35.0	53.4%	38.1	90.3%

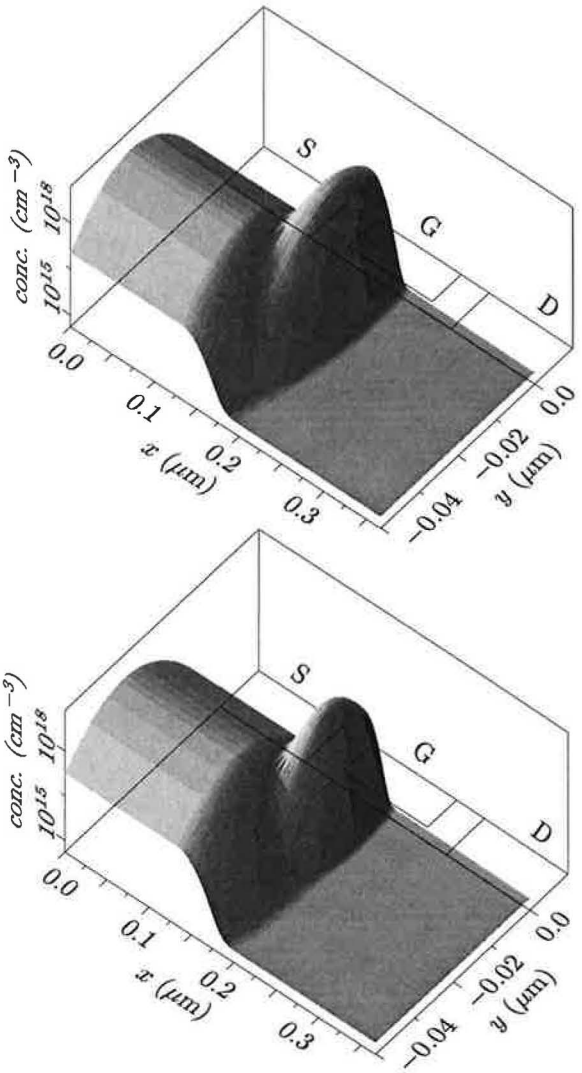

Figure 9: The gate delay time optimization results using Gaussian functions for the NMOS (top) and PMOS (bottom) transistors with 0.1 μm gate lengths

It is to note that the performance gains due to gate delay time optimizations are calculated from the inverse of the average delay time t_d which is a metric for the speed of the inverter. They turn out to be higher than one would have predicted from the results of drive current optimizations. Obviously, the improved drive current is not the only reason for the increased inverter speed. The decreased device capacitances, mainly the drain-bulk junction capacitance, also contribute to the enhanced inverter characteristics. The lighter background doping of the optimized devices provides a much smaller drain-bulk capacitance compared to the uniformly doped devices which reduces the total capacitance at the output node of the inverter and, therefore, the delay time.

To verify that the resulting delay times using the single stage inverter model of Fig. 5 are realistic values, a five stage ring oscillator circuit is simulated using mixed-mode transient simulations. This rather complex simulation task can be performed thanks to the rigorous mixed-mode simulation capabilities of MINIMOS-NT [5].

The node voltages of the ring oscillator using uniformly doped devices and devices with optimized doping profiles are depicted in Fig. 10 for the 0.1 μm device generation.

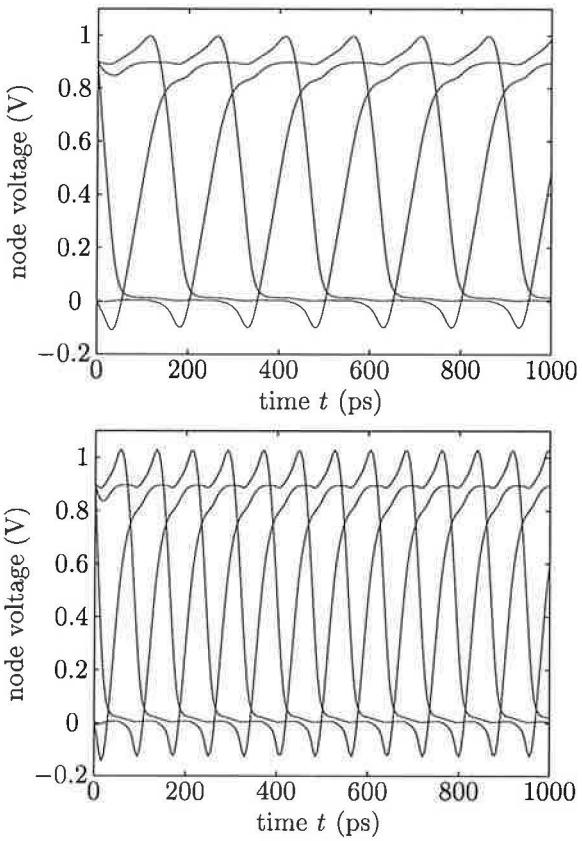

Figure 10: The node voltages of a five stage ring oscillator using uniformly doped devices (top) and devices with optimized doping profiles (bottom)

Table 4 compares the calculated ring oscillator gate delay times with the results obtained from the infinite inverter chain emulations using the single stage model. The ring oscillator delay times are slightly higher than the single stage inverter delay times because the overshoot in the output voltage is neglected when it is used as a new input voltage in the inverter chain emulation. However, the single stage inverter model has proven to be a very good approximation for the realistic case that occurs in a digital circuit. For optimization purposes a correct qualitative behavior of a model is the primary concern because the goal is to improve a certain performance metric.

Table 4: Comparison of the gate delay times of the ring oscillator to the single stage inverter model

	Generation A		Generation B	
	ring osci.	inverter	ring osci.	inverter
uniform	55.3 ps	53.7 ps	74.3 ps	72.5 ps
two-dim.	35.4 ps	34.9 ps	38.9 ps	36.8 ps

6 CONCLUSION

It has been shown that automatic closed-loop optimizations with TCAD frameworks offer a great potential for high-level device design applications even for complex performance goals and a large number of design parameters. Due to its flexibility, the presented design optimization procedure can be applied to many different optimization tasks, not only for semiconductor applications, but in all fields of science where numerical simulation can be used to predict a system's performance.

ACKNOWLEDGMENT

The authors would like to thank Andreas Wild for many fruitful discussions. This work was financially supported by Intel, Inc.

REFERENCES

[1] R. Plasun, M. Stockinger, and S. Selberherr, "Integrated Optimization Capabilities in the VISTA Technology CAD Framework," *IEEE Trans.Computer-Aided Design of Integrated Circuits and Systems*, vol. 17, no. 12, pp. 1244–1251, 1998.

[2] R. Strasser, *Rigorous TCAD Investigations on Semiconductor Fabrication Technology*, Dissertation, Technische Universität Wien, 1999.

[3] R. Plasun, *Optimization of VLSI Semiconductor Devices*, Dissertation, Technische Universität Wien, 1999.

[4] R. Strasser and S. Selberherr, "Parallel and Distributed TCAD Simulations Using Dynamic Load Balancing," *Proc. Simulation of Semiconductor Processes and Devices*, pp. 89–92, Leuven, Belgium, Sept. 1998.

[5] T. Grasser, V. Palankovski, G. Schrom, and S. Selberherr, "Hydrodynamic Mixed-Mode Simulation," *Proc. Simulation of Semiconductor Processes and Devices*, pp. 247–250, Leuven, Belgium, Sept. 1998.

[6] M. Stockinger, A. Wild, and S. Selberherr, "Drive Performance of an Asymmetric MOSFET Structure: The Peak Device," *Microelectronics Journal*, vol. 30, no. 3, pp. 229–233, 1999.

[7] C.-C. Shen, J. Murguia, N. Goldsman, M. Peckerar, J. Melngailis, and D.A. Antoniadis, "Use of Focused-Ion-Beam and Modeling to Optimize Submicron MOSFET Characteristics," *IEEE Trans.Electron Devices*, vol. 45, no. 2, pp. 453–459, 1998.