Sub-wavelength Lithography and Variability Aware SRAM Characterization

Petr DOBROVOLNÝ, Miguel MIRANDA, Paul ZUBER

Dept. of Design and Tech. Enablement, Div. of Process Technology, Imec vzw., Kapeldreef 75, B-3001 Leuven, Belgium

dobrovol@imec.be, miranda@imec.be, zuberp@imec.be

Abstract. *With shrinking of minimum feature size of advanced technology nodes, the impact of litho process variations on the resulting electrical parameters of printed circuits dramatically increases. Litho process variations correspond to random changes in the actual optical conditions (dose and focus) which develop at every mask exposure, hence from die to die. In this way the litho process variations act as a global variability component affecting all devices on a particular die in the same way. In contrast to this, the intrinsic variability of the devices and interconnects originating mostly from local Random Dopant Fluctuations (RDF) and Line Edge Roughness (LER) has a purely spatially uncorrelated component. Yet, it is not clear which of the two limits scaling down variability sensitive circuits such as SRAM beyond 45 nm. This paper presents a tool flow to perform SRAM wide statistical analysis subject to combinations of global litho and local variability components. The tool flow is illustrated in 45 nm industry grade SRAM vehicle. Selected case studies show how this tool flow successfully captures non-trivial statistical interactions between the SRAM cell and the periphery, otherwise less visible when using statistical electrical simulations of the critical path alone.*

Keywords

Litho process and technology variability, statistical SRAM analysis, yield prediction.

1. Introduction

In this paper we present an approach offering full memory statistical analysis capabilities addressing litho and technology process variation effects and aiming at improving design productivity of embedded SRAM products.

We show how the most likely reasons for statistical failure can be anticipated at design time so as to correct weak design spots before tape-out, hence avoiding costly silicon spin iterations. The technique provides key help to memory and system designers to estimate parametric and functional yield loss due to statistical parametric spreads in threshold voltage and/or current gain of the devices. Hence it is of especial value to the design of embedded SRAMs, which are considered to be the most sensitive SoC components to process variations.

The estimated yield loss can be of parametric nature (i.e., failure of the memory access to meet the target cycle time or insufficient read margin for successful operation or similar); but it can be of functional nature (i.e., failure to meet a stability criteria or any other pass/fail functional check).

The strength of the approach lies on successfully capturing all (non-trivial) memory-wide statistical interactions between the SRAM cell and the periphery, otherwise less visible when using statistical electrical simulations of the critical path alone.

The tool flow requires five main input items. The first is a transistor level netlist description of a segment of the memory describing all circuitry involved from input to output. The second one is a memory layout designed according to prescribed design rules. The third one is information about litho process conditions. For the needs of litho variability analysis a process conditions region is represented by a set of pairs – dose/focus numbers – carefully selected from the border of a litho process windows. The fourth one is a set of parameters describing the internal architecture of the memory, thus how the memory is built from the segment information, including redundancy and error correction code infrastructure. The fifth one is information about the variability of the devices and interconnects used in the underlying technology. This information can be provided in either the form of statistical distributions of certain transistor parameters, scattered data obtained via statistical simulation of the device or just plain DC current-voltage statistical relationships of fabricated devices obtained during silicon characterization.

The approach consists of two main phases. The first phase (see Fig. 1) performs litho variability analysis of a memory layout. Based on litho process analysis, the critical path netlist of a memory is updated, namely transistor sizes are adjusted. The degree of litho process impact on device size changes depends on a particular process conditions and on the layout itself, specifically on design rules applied for memory layout design. Thus the

output of the first phase is the set of updated memory critical path netlists. These netlists represent design corners for a given litho process conditions and chosen design rules.

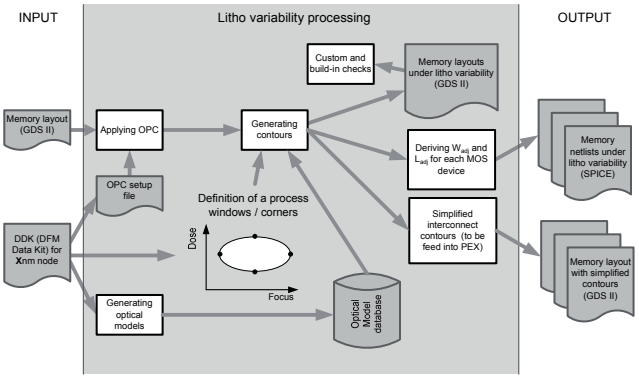

Fig. 1. Litho variability analysis flow.

The second phase (see Fig. 2) consists of two main steps. The first step performs a statistical electrical simulation of the full SRAM critical path netlist. Such critical path includes not only bitline read/write circuits, but also the rest of the main SRAM circuit blocks (hereafter called memory 'islands'). Examples are the row decoding, timing circuit and output stage buffers (a more formal definition of a memory island can be found in Section 3). During this phase, statically correlated parametric data and pass/fail information obtained from the critical path simulations is thus collected via inserted measurement and check point statements in the netlist.

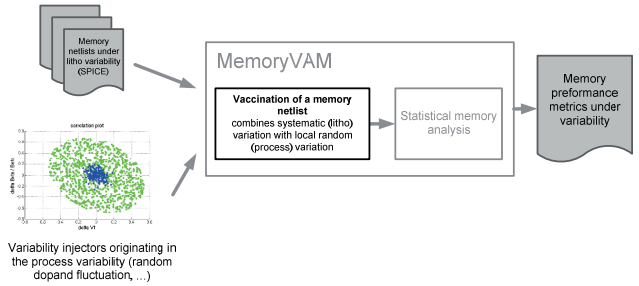

Fig. 2. Statistical memory characterization.

The key to this strategy is the ability to complement the analysis of a nominal memory model under test with statistically sampled variant of the devices. An in-house developed enhanced Monte Carlo technique is used to significantly reduce the amount of statistical simulations needed to achieve a particular level of confidence [1]. However any other existing technique to enhance statistical sampling could be used as well [2], [3]. At this stage, sensitivity information of the impact of process variations resulting from the different SRAM islands in the targeted measurement point is collected as well.

During the second step, the collected statistical data and sensitivity information is used to reconstruct the full memory parametric and pass/fail behavior. This is done by using a mix of statistical and algorithmic data manipulation techniques. This phase is rather fast, several tens megabyte memory can be analyzed in few seconds. Based on this, statistical information on the critical path percolates to the complete SRAM organization level, resulting in a realistic prediction of the yield as perceived by the memory tester and/or equivalent BIST (built-in-self-testing) technique.

2. Related Work

Emerging statistical techniques such as statistical static timing analysis (SSTA) or statistical circuit level simulation tools have been proposed to avoid conventional process corner analysis when handling random process variations. Instruments to predict local process variations of circuits comprise Monte Carlo like runs around a nominal simulation for transistor level netlists or adding probabilistic awareness to e.g., timing analysis algorithms at the verilog level. This adds a parametric yield dimension to the verification flow before tapeout, allowing the designer to take decisions and make changes where/when (s)he still can.

For memories, virtually no commercially available solution exists however. Several issues make memories especially challenging:

- Usually the nominal simulation of a full memory is reduced to the critical path netlist, assuming every other path behaves the same. However this approach particularly fails under local process variations where device to device uncorrelated variability makes every bitcell access operation behave differently. Since a memory is as good as its worst path, the memory statistics for instance of the read margin or access time is the distribution of the worst of all its cells. As a result, simulating the critical path netlist under variability does not model the full memory statistics correctly.
- Works considering the bitcell alone without its periphery [4] manage to reduce the sample sizes and transistor counts effectively, but also entails incomplete analysis. For instance the different memory components influencing the read operation of the cell can affect its read margin. Indeed, variations can affect the sense amplifier offset and the timing circuit that controls its activation, the row decoder that enables the word line activation, and especially the cell's capability to discharge the bitline. Accounting for the worst case situation of each of these effects would lead to pessimistic estimations of the read margin. As a con-sequence, we must simulate the entire equivalent circuit's operation under variability to obtain realistic results.
- On top of that, attention must be paid to architectural correlations of bitcells and sense amplifiers and other memory parts. A worst case cell instance is not necessarily in the same path with the worst case sense

amplifier or the worst case row driver logic so that a blind worst case combination would lead to over pessimistic results again. In order to get the bitcell statistics, a pure Monte Carlo based approach is unable to give results at all in higher yield realms. It simply fails to capture the tails of the real distribution for a reasonable amount of simu-lations. Moreover, the higher the instance count of a certain memory island circuit, the more tail exploration is required. Importance sampling variants have been used in literature in times [5], [2], [6], [7] in order to derive tail statistics of a single cell and to predict a memory failure probability.

A generic analytical approach to predict the statistics of a system from the given probability density function of its subsystem components appeared in [8], and particularly for memories in [9]. However, both methods fail to accurately capture all the interactions between the different memory islands. The two typical phenomena resulting from these extreme value problems are highly skewed and shifted Probability Density Functions (PDF). This is why simply using a statistically enhanced (importance sampling) engine on the critical path alone or even a more systematic Design of Experiments [10] approach to gather the mismatch statistics of the critical path netlist would fail as well.

In this work, we consider all the above considerations and take them one step further. We report a method and its implementation in a tool flow hereafter called Litho Aware Memory Variability Aware Modeling or Litho-MemoryVAM in short – based on technique that predicts the correct memory wide statistics of any parameter that can be measured in a SPICE/SPECTRE testbench, such as access time, power, stability checks such as read voltage, and so on. The method lies on a mix of sensitivity analysis of the different memory 'islands' conforming the critical path of the memory to process variations. Thus, such sensitivity analysis is done at a larger granularity than the transistor level proposed so far for analog circuits [11], hence leading to a more efficient amount of simulation runs needed hence much less CPU time.

3. Statistical SRAM Analysis

In this section we describe how the full memory statistics can be recovered by combining island statistics in a specific way described in the following.

3.1 Litho Variability Analysis

With shrinking of minimum feature size of advanced technology nodes, the impact of litho process on resulting electrical parameters of printed circuits dramatically increases. The final printed and etched shapes in silicon are quite different from original drawn rectangular shapes (see Fig. 3).

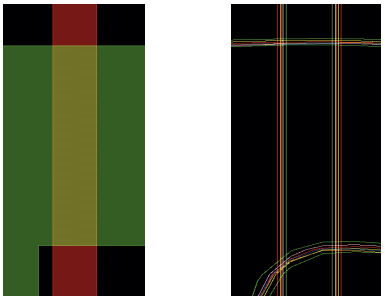

Fig. 3. Example of a drawn rectangular shape and corresponding final shape printed in silicon.

For the needs of post-lithography circuit simulation, the nonrectangular gate shapes have to be analyzed and modeled. Several analysis and modeling approaches were suggested [12], [13]. Solutions based on TCAD simulations usually lead to the replacement of original nonrectangular gate device by a set of parallel rectangular devices with carefully defined model parameters (see Fig. 4).

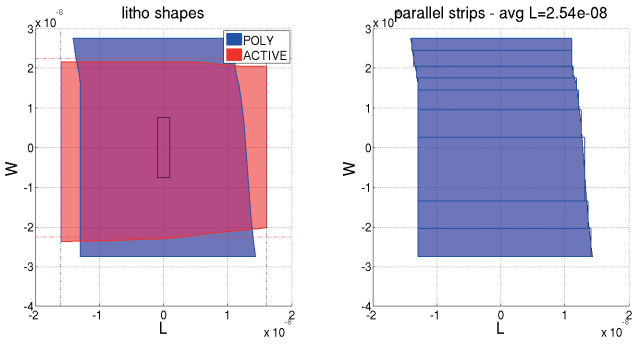

Fig. 4. Horizontal segmentation of nonrectangular poly gate layer shape resulting in linear superposition of currents.

Another branch of solutions relies on the original Spice compact model but derives adjusted sizes W_{adj} and L_{adj} replacing the original drawn sizes W and L (see Fig. 5). This approach is fully compatible with BSIM 4.5 compact model definition and provides enough accuracy for 40 nm and 32 nm technology node [14].

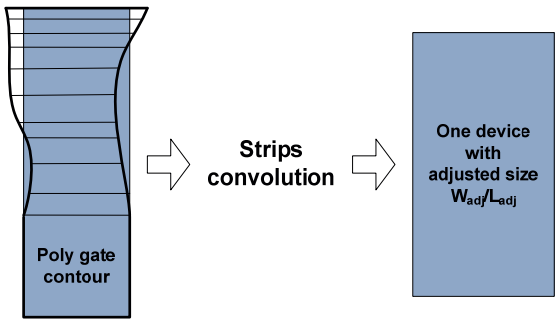

Fig. 5. Deriving adjusted size of a device based on strips convolution of its printed nonrectangular shape.

Taking into account the variability of litho process and context dependency, the shape distortion problem is going to be even more complex and subject to statistical variations. The variability of litho process corresponds to

variability in actual litho conditions – described by actual dose and focus values - and varies from die to die. In this way the litho process variability acts as a systematic variability component affecting all devices on a particular die in the same way. Moreover the impact of applied litho process also depends on the actual layout context. It means that identical devices printed in different locations will actually print differently, and hence behave as if they were two different devices. At the level of layout design, the context dependency is defined by design rules and available layout options (e.g. availability of local interconnect).

Our tool flow relies on Model Based Extraction (MBE) approach developed by Mentor Graphics and implemented in Calibre Litho-Friendly Design (LFD) tool. It accepts as an input a memory netlist and a memory layout designed under particular design rules. We utilize Calibre LFD for litho process simulation under defined litho process conditions. For a set of carefully selected dose-focus pairs from defined process window, the set of updated memory netlists is generated. The impact of litho process is introduced through adjusted sizes of each transistor device. Thus the set of updated memory netlist represent design corners defined by litho process variability and applied design rules.

In the subsequent stage, memory netlists with introduced litho variability are subject of applying variability injectors representing influence of process technology variability.

3.2 Technology Process Variability Injection

In our experiments, we extracted populations of ΔV_{th} and $\Delta\beta$ pairs for different MOSFET types from industry grade statistical device compact model library. Comparison of our two-parameter model to the reference using 10,000 Monte Carlo runs shows excellent agreement of mean, spread and correlation of circuit performance metrics. This is not true however for those models assuming V_{th} alone as source of random process variations such as random dopant fluctuations or alike. This can cause some 20 % difference in standard deviation of important metrics like gate leakage or delay.

Since Spice itself does not support importance sampling or any other form of statistical enhancement technique, we introduce the concept of circuit variability injection. Variability injection denotes the process of transforming a spice level netlist with any number and type of transistors into N variants of the same. These variants differ from the original netlist in that variability injection sources are added to the MOSFET instances of an island. A voltage source at the gate and a drain-current controlled current source along source-drain are used to model ΔV_{th} and $\Delta\beta$, respectively. The values of these injection sources differ among the variants and obey the underlying distribution. W and L are processed to comply with Pelgrom's rule [15].

3.3 Critical Path Analysis

We propose to generate the sensitivity of critical path parameters to process variations in different islands. An island is formally defined as a set of transistors forming a memory component whose instance count (multiplicity) and connectivity differs to other islands. Our motivation is three-fold:

- Only a simulation of the entire critical path netlist allows to take any parameter measurement in the original SPICE netlist, especially those which 'cross' island boundaries or are influenced by several islands. Assembling such parameters from smaller circuit measures (or equivalents) would be far less straightforward if not impossible.
- The different islands occur once along the critical path netlist but in a full memory, they occur with different multiplicities. Sensitivities offer a way to account for real instance counts, as will be shown in Section 3.4.
- Considering variability in subsets of all transistors, increases the effectiveness of any importance sampling technique [3].

3.4 Architecture Aware Scaling

After simulating the critical path sensitivity statistics for all islands, the final important step is deriving the memory statistics from the sensitivities. Since this happens under awareness of the connectivity of the islands, we refer to this step as Architecture Aware Scaling (AAS). It consists of the following consecutive steps.

- For every island $i = 1 \ldots I$, pick an M_{i}-sized sample from the sensitivity distribution ΔP_{i} of the memory due to island i. The parameter P can be multi-valued and contain any measure that was taken in the original spice netlist. ΔP is the variation of P shifted around nominal.
- Systematically list all M possible paths through a memory with all its island instances along them. There will be paths sharing sense amplifiers, others sharing row logic. All bitcells are in unique paths, and the timing generator is shared by all paths. Any path will thus be a particular combination of island indexes: $p_{\mathrm{j}} = (z_1, \ldots z_I)$, $j = 1 \ldots M$, $z_i \in \{1 \ldots M_i\}$. This connects the ΔP_{i} statistics according to the topology inside the memory.
- Re-assemble path parameters from sensitivities of islands that form the path. For every path, $P_j \approx P_0 + \Sigma_{\mathrm{i}} \Delta P_i(z_i)$ where P_0 is the nominal point. This rule provided good accuracy ($\leq$ 1 % RMS error) on all examples used in this work. Optionally, a position dependent component $\Delta P_{\mathrm{pos}}(x, y)$ can be added to

reflect systematic and random variations within a correlation length.

- Select the worst of the *M* paths and assign that value to the memory observation. Depending on the meaning of P this can be the max operator (e.g. late mode timing), the min operator (e.g. read voltage, early mode timing) or the average for power.
- Repeat *n* times with different random sequence to generate memory statistics.

4. Results and Applications

Litho-MemoryVAM connects the different steps described above using MATLAB on top of SPICE. It can be configured with any number of islands and hierarchy depth, such as banks, local word lines, multiplexed output buffers and the like.

4.1 Sensitivity Analysis

In Section 3.3 we proposed to analyze the sensitivity of critical path metrics to process variations in particular islands (blocks of transistors) for use as an input to architecture aware scaling, i.e. as intermediate results. As it turns out such intermediate results are per se already providing valuable information to the designer. Fig. 6 shows such results for a 250 kBit memory block in 45 nm.

We selected the timing circuitry, the sense amplifier block and the bitcell islands, and analyzed the effect of variations therein to the access time and read margin. The user can easily see for example that for improving speed the best places to optimize are the periphery circuitry and the sense amplifier blocks, while bitcells have little influence. The opposite is true for the read margin. Similar graphs can be plotted for any other measure as well.

4.2 Critical Path vs. Full Memory Analysis

One attempt to derive the worst case behavior of a memory might be a corner simulation of the critical path netlist. Another attempt might be to simulate the critical path netlist statistically. We show such results in Fig. 7 for one an industry grade SRAM block and compare it to Litho-MemoryVAM.

To stress the effect caused by local random variations we have separated the graph into local and global variations. It can be seen that even though the global variations (still) cause higher uncertainties, it is the local random variability that causes an additional shift of the cloud away from the nominal point.

It is interesting to note that the global variations of access time and read voltage are negatively correlated, while the local random variations of the memory are positively correlated. This effect is caused by a combination of maximum and minimum operators, on the two respective parameters.

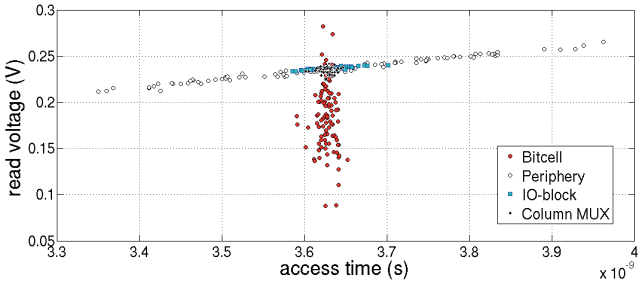

Fig. 6. Sensitivity analysis results for memory islands.

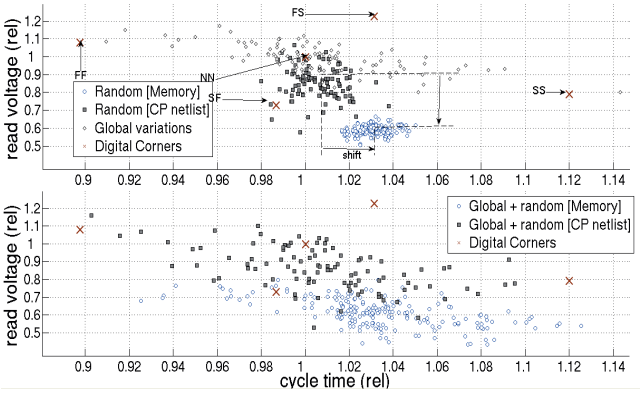

Fig. 7. Every single dot in the scatter plot represents one observation of a critical path netlist or of a full memory. Digital corners fail to characterize the true memory worst case.

Such observations cannot be made by just statistically simulating the critical path netlist, even if the approach is to simulate local variations around a global corner. Combined (global & local) corners would need on top of the global component also an equivalent local random margin to capture the architecture-dependent shift. Traditional digital corners are clearly indisputable for memory statistics.

5. Conclusions and Future Work

In this paper we defined an automated approach to enable statistical timing, power and yield analysis for full SRAM arrays. Case studies based on industry-grade embedded SRAM in 45 nm technology nodes are used to illustrate the approach. We show quantitatively how our method can be used to avoid costly design iterations.

All presented results were generated under optimal litho conditions - optimal dose and focus values for underlying litho process. The future work will bring comparison of analysis results arising from different design corners created by lithography (systematic) variation with applied random variations coming from technology process variation.

Acknowledgements

The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under the grant agreement n° 248538 (SYNAPTIC project).

References

[1] ZUBER, P., MATVEJEV, V., DOBROVOLNY, P., ROUSSEL, P., MIRANDA, M. Using exponent Monte Carlo for quick statistical circuit simulation. In *Proceedings of the International Workshop on Power and Timing Modeling, Optimization and Simulation PATMOS 09*. Delft (Netherlands), 2009, p. 36 – 45.

[2] KANJ, R., JOSHI, S., NASSIF, R. Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare failure events. In *Proceedings of the 43rd Annual Design Automation Conference DAC 2006*. New York (NY, USA), 2006, p. 69 - 72.

[3] HOCEVAR, D., LIGHTNER, M., TRICK, T. A study of variance reduction techniques for estimating circuit yields. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 1983, vol. 2, no. 3, p. 180 - 192.

[4] ZHOU, Y. *et al*. The impact of beol lithography effects on the SRAM cell performance and yield. In *Proceedings of the 2009 International Symposium on Quality Electronic Design ISQED 09.* New York (NY, USA), 2009, p. 607 - 612.

[5] NHO, H., YOON, S., WONG, JUNG, S. Numerical estimation of yield in sub-100-nm SRAM design using Monte Carlo simulation. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2008, vol. 55, no. 9, p. 907 - 911.

[6] CHEN, G. K., BLAAUW, D., MUDGE, T., SYLVESTER, D. Yield driven near-threshold SRAM design. In *IEEE/ACM International Conference in Computer-Aided Design ICCAD 2007.* San Jose (CA, USA), 2007, p. 660 - 666.

[7] DOLECEK, L., QAZI, M., SHAH, D., CHANDRAKASAN, A. Breaking the simulation barrier: SRAM evaluation through norm minimization. In *IEEE/ACM International Conference on Computer-Aided Design ICCAD 2008.* San Jose (CA, USA), 2008, p. 322 - 329.

[8] MIRANDA, M., DIERICKX, B., ZUBER, P., DOBROVOLNÝ, P., KUTSCHERAUER, F., ROUSSEL, P., POLIAKOV, P. Variability aware modeling of SoCs: from device variations to manufactured system yield. In *Proceedings of the 2009 International Symposium on Quality Electronic Design ISQED09.* San Jose (CA, USA), 2009, p. 547 - 553.

[9] AITKEN, R., IDGUNJI, S. Worst-case design and margin for embedded SRAM. In *Proceedings of the Conference Design, Automation & Test in Europe*. Nice (France), 2007, p. 1 - 6.

[10] MYERS, R. H., MONTGOMERY, D. C. *Response Surface Methodology: Process and Product in Optimization Using Designed Experiments*, 2nd ed. New York (NY, USA): John Wiley & Sons, 2002.

[11] MCCONAGHY, T., GIELEN, G. Globally reliable variation-aware sizing of analog integrated circuits via response surfaces and structural homotopy. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* 2009, vol. 28, no. 11, p. 1627 - 1640.

[12] SEON-DONG KIM, WADA, H., WOO, J. C. S. TCAD-based statistical analysis and modeling of gate line-edge roughness effect on nanoscale MOS transistor performance and scaling. *IEEE Transactions on Semiconductor Manufacturing*, 2004, vol. 17, no. 2.

[13] SINGHA, R., BALIJEPALLI, A., SUBRAMANIAM, A., LIU, F., NASSIF, S. Modeling and analysis of non-rectangular gate for post-lithography circuit simulation. In *Proceedings of the 44th Annual Design Automation Conference DAC 2007*. San Diego (CA, USA), 2007, p. 823 - 828.

[14] Mentor Graphics Corporation. *Calibre Litho-Friendly Design User's Manual*, 2008.

[15] PELGROM, M. J. M., DUINMAIJER, A. C. J., WELBERS, A. P. G. Matching properties of MOS transistors. *IEEE Journal of Solid-State Circuits,* 1989, vol. 24, no. 5, p. 1433 - 1439.

About Authors...

Petr DOBROVOLNÝ received his M.Sc. and Ph.D. degrees from the Brno University of Technology, Dept. of Microelectronics, in 1987 and 1998, respectively. During his Ph.D. studies he investigated the problem of the symbolic analysis of large analog circuits in the cooperation with the KUL, Dept. ESAT-MICAS. Since 1999, he has been with imec, where he is currently a senior research in the VAM (Variability Aware Modeling) team. He focuses on the research and development of CAD tools for circuit characterization under process variability and in the presence of time dependent degradation mechanisms.

Miguel MIRANDA is Principal Scientist in the Design & Technology Enablement Department at imec (Belgium), where he functions as System Technology Architect and leads the R&D engineering efforts of the VAM (or Variability Aware Modeling) Project; a Design-for-Manufacturability tool flow providing statistical characterization at every level of abstraction of the electronic design hierarchy, from library characterization to processor core characterization, including the world's first memory wide variability analysis tool. Miguel holds more than 10 European and USA patents and he is (co)author of more than 120 peer-reviewed publications in international conferences and journals. He has served in various Technical Program and Organizing Committees of international conferences (DATE, DAC, ESWEEK, ESTIMEDIA, CODES+ISSS, IOLTS, ...) and served as associated editor of several international journals.

Paul ZUBER has a Ph.D. from the Tech. Univ. of Munich since 2007 and has proven experience in DFV topics, with recent key papers in design and technology conferences (DAC10,11, DATE10,11, ISQED09, VLSI Techno Symp.11, etc). Dr Zuber is (co)author of 2 patents in circuit/system variability analysis and was imec's lead researcher in one of the 1st European initiatives in DFV (the FP7 REALITY project).