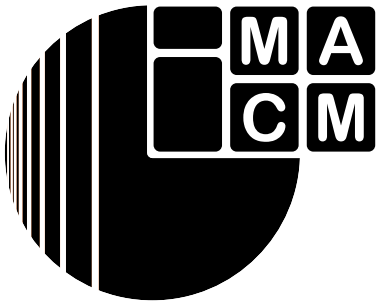

Bergische Universität Wuppertal

Fachbereich Mathematik und Naturwissenschaften

Institute of Mathematical Modelling, Analysis and Computational Mathematics (IMACM)

Preprint BUW-IMACM 12/30

Toby Doorn, Jan ter Maten, Alessandro Di Bucchianico, Theo Beelen, Rick Janssen

Access Time Optimization of SRAM Memory with Statistical Yield Constraint

November 2012

http://www.math.uni-wuppertal.de

Access Time Optimization of SRAM Memory with Statistical Yield Constraint

Toby DOORN[1], *Jan ter MATEN*[2,3], *Alessandro DI BUCCHIANICO*[2], *Theo BEELEN*[1], *Rick JANSSEN*[1]

[1] NXP Semiconductors, High Tech Campus 32 and 46, 5656 AE Eindhoven, the Netherlands
[2] Dept. Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, the Netherlands
[3] Chair of Applied Mathematics / Numerical Analysis, Fachbereich C, Bergische Universität Wuppertal, Gaußstraße 20, D-42119 Wuppertal, Germany

{Toby.Doorn,Theo.G.J.Beelen,Rick.Janssen}@nxp.com, {E.J.W.ter.Maten,A.D.Bucchianico}@tue.nl, Jan.ter.Maten@math.uni-wuppertal.de

Abstract. *A product may fail when design parameters are subject to large deviations. To guarantee yield one likes to determine bounds on the parameter range such that the fail probability P_{fail} is small. For Static Random Access Memory (SRAM) characteristics like Static Noise Margin and Read Current, obtained from simulation output, are important in the failure criteria. They also have non-Gaussian distributions. With regular Monte Carlo (MC) sampling we can simply determine the fraction of failures when varying parameters. We are interested to efficiently sample for a tiny fail probability $P_{\text{fail}} \leq 10^{-10}$. For a normal distribution this corresponds with parameter variations up to 6.4 times the standard deviation σ. Importance Sampling (IS) allows to tune Monte Carlo sampling to areas of particular interest while correcting the counting of failure events with a correction factor. To estimate the number of samples needed we apply Large Deviations Theory, first to sharply estimate the amount of samples needed for regular MC, and next for IS. With a suitably chosen distribution IS can be orders more efficient than regular MC to determine the fail probability P_{fail}. We apply this to determine the fail probabilities the SRAM characteristics Static Noise Margin and Read Current. Next we accurately and efficiently minimize the access time of an SRAM block, consisting of SRAM cells and a (selecting) Sense Amplifier, while guaranteeing a statistical constraint on the yield target.*

Keywords

Importance sampling, monte carlo, large deviations, failure probabilities

1. Introduction

As transistor dimensions of Static Random Access Memory (SRAM) become smaller with each new technology generation, they become increasingly susceptible to statistical variations in their parameters. These statistical variations may result in failing memory. An SRAM is used as a building block for the construction of large Integrated Circuits (ICs), providing Megabits of memory. To ensure that a digital bit cell in SRAM does not degrade the yield (fraction of functional devices) of ICs, very small failure probabilities are necessary [2]. For instance, in SRAM memory design one aims to get less than 0.1% yield loss for a 10Mbit memory, which means that at most 1 in 10 billion cells fails ($P_{\text{fail}} \leq 10^{-10}$; for a one-sided tail probability this corresponds with a -6.4σ parameter variation when dealing with a normal distribution; here σ is the standard variation). To simulate this, regular Monte-Carlo (MC) requires a huge number of simulations, despite some speed-up techniques that are available in commercial simulation tools (Latin Hypercube, stratification, quasi Monte Carlo, etc.). Importance Sampling (IS) [1] is a sampling technique that is relatively easy to implement. Practice shows that by IS one can obtain sufficiently accurate results in a much more efficient way than by MC [3,5,6,9]. Also some variants show up [4]. Section 2 and 3 provide sharp upper bounds for the number of samples needed by MC and by IS more advanced technique that provides sufficiently accurate results and is relatively easy to implement. A speed up of several orders can be achieved when compared to regular Monte Carlo methods.

2. Regular Monte Carlo

Let Y be a real-valued random variable with probability density function f. We assume that N independent random observations Y_i ($i = 1,\ldots,N$) of Y are taken and define, for a given set $A = (-\infty, x)$, the event indicator $X_i = I_A(Y_i)$, where $I_A(Y_i) = 1$ if $Y_i \in A$ and 0 otherwise. Then $p_f^{\text{MC}}(A) = \frac{1}{N}\sum_{i=1}^{N} X_i$ estimates $p = \int_{-\infty}^{x} f(z)dz = P(Y \in A)$. The X_i are Bernoulli distributed, hence $Np_f^{\text{MC}} \sim \text{Bin}(N,p)$ is Binomially distributed (N samples, each with success probability p), and thus for the expectation one has $\text{E}(p_f^{\text{MC}}) = \frac{1}{N}Np = p$, and for the variance $\text{Var}(p_f^{\text{MC}}) \equiv \sigma^2(p_f^{\text{MC}}) = \frac{p(1-p)}{N}$. Here

$\sigma(p_f^{\rm MC})$ is the corresponding standard deviation.

Let $\Phi(x) = \int_{-\infty}^{x} e^{-z^2/2} dz$ be the cumulative probability function of the normal density function and define z_α by $\Phi(-z_\alpha) = \alpha$, see Fig. 1 for an impression. For $N_{\rm MC}$ large

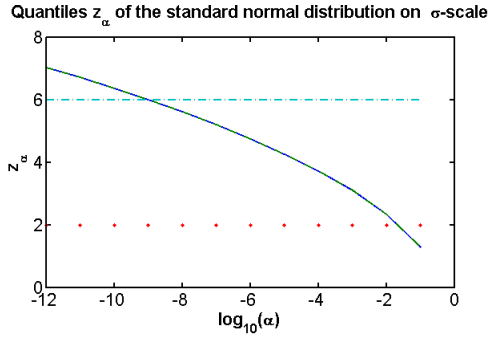

Fig. 1. Powers of tail accuracy, $\log_{10}(\alpha)$, versus quantiles z_α of the normal distribution along σ-scale. Our interest goes to variations up to 6σ.

enough we can apply the Central Limit Theorem (CLT) and derive

$$P(|p_f^{\rm MC} - p| > \varepsilon) \quad = \quad P(\frac{|p_f^{\rm MC} - p|}{\sigma(p_f^{\rm MC})} > z)$$
$$\stackrel{N_{\rm MC}\to\infty}{\longrightarrow} \quad 2\Phi(-z) \le 2\Phi(-z_{\alpha/2}) = \alpha,$$

where $z = \varepsilon/\sqrt{p(1-p)/N_{\rm MC}}$ and $N_{\rm MC} = N$. Hence, if $z \ge z_{\alpha/2}$ we deduce

$$N_{\rm MC} \ge p(1-p)\left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 = \frac{1-p}{p}\left(\frac{z_{\alpha/2}}{\nu}\right)^2, \qquad (1)$$

for $\varepsilon = \nu p$. Here we assume $\nu = 0.1$ and $p = 10^{-10}$. Now let $\alpha = 0.02$, then $z_{\alpha/2} \approx 2$. Then (1) implies $N_{\rm MC} \ge 4\ 10^{12}$. If we do not know p, we can use $p(1-p) \ge 1/4$, yielding $N_{\rm MC} \ge \frac{1}{4}\left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 = 10^{22}$. And if $N_{\rm MC}$ is not large enough to apply the CLT, Chebyshev's inequality even results to $N_{\rm MC} \ge 10^{24}$. These general bounds are much too pessimistic. Large Deviations Theory (LDT) [1,7] results in a sharp upper bound that nicely involves $N_{\rm MC}$

$$P(|p_f^{\rm MC} - p| > \nu p) \quad \le \quad \exp\left(-\frac{N_{\rm MC}}{2}\frac{p}{1-p}\nu^2\right), \quad (2)$$

for all $N_{\rm MC}$, with a possible exception of finitely many. For a proof, see [11, 12]. The exponential type of bound in (2) is also valid from below and thus is sharp. For $\nu = 0.1$, $p = 10^{-10}$ and $\alpha = 0.02$, as above, we find: $N_{\rm MC} \ge 8\ 10^{12}$ (which is thus a sharp result). Note that an extra k-th decimal in ν increases $N_{\rm MC}$ with a factor k^2.

3. Importance Sampling

With Importance Sampling we sample the Y_i according to a different distribution function g and observe that $p_f(A) = \int_{-\infty}^{x} f(z)dz = \int_{-\infty}^{x} \frac{f(z)}{g(z)} g(z)dz$. We define a weighted success indicator $V = V(A) = I_A(Y)f(Y)/g(Y)$. Then with the g-distribution we have for the expectation $\mathrm{E}_g(V) = \int I_A(y)\frac{f(y)}{g(y)}g(y)dy = \int_{-\infty}^{x} f(y)dy = p_f(A)$. Hence if we determine $V_i = I_A(Y_i)f(Y_i)/g(Y_i)$ from g-distributed Y_i we can define $p_g^{\rm IS}(A) = \frac{1}{N}\sum_{i=1}^{N} V_i$. Its expectation becomes $\mathrm{E}_g\left(p_g^{\rm IS}\right) = \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}_g\left(V_i\right) = p_f(A)$. When also $\frac{f(z)}{g(z)} \le 1$ on A we derive after some calculation $\mathrm{Var}_g\left(p_g^{\rm IS}\right) \le \mathrm{Var}_f\left(p_f^{\rm MC}\right)$ (variance reduction, using the same number of samples). This does not yet imply more efficiency. However, similar to (2), we derive (in which $N_{\rm IS} = N$), for $N_{\rm IS}$ large enough

$$P\left(|p_g^{\rm IS} - p| > \nu p\right) \quad \le \quad \exp\left(-\frac{N_{\rm IS}\ p^2}{2\mathrm{Var}_g(V)}\nu^2\right). \quad (3)$$

For a proof, we again refer to [11, 12]. Assuming the same upper bounds values in (2) and (3), comparing them gives $\frac{N_{IS}}{N_{MC}} = \frac{\mathrm{Var}_g(V)}{p(1-p)} = \frac{\mathrm{E}_g(V^2)-p^2}{p(1-p)}$. Now, suppose $p \le \kappa$ and

$$\frac{f(z)}{g(z)} \le \kappa < 1, \ \text{ on } A. \qquad (4)$$

Then, with $q = 1-p$, we obtain

$$\frac{N_{IS}}{N_{MC}} = \frac{\mathrm{E}_g(V^2)}{pq} - \frac{p}{q} \le \frac{\kappa}{q} - \frac{p}{q} \le \kappa(1+\zeta) \qquad (5)$$

for $|(1-\frac{1}{\kappa})p + \mathcal{O}(p^2)| \le \zeta$, which for $\kappa = 0.1$ and $p = 10^{-10}$ means that $\zeta \le 10^{-9}$. Hence, for $\kappa = 0.1$, we can take an order less samples with Importance Sampling to get the same accuracy as with regular Monte Carlo. This even becomes better with smaller κ. By Importance Sampling we gain efficiency; this is the main message. Also the asymptotic accuracy improves when compared to regular Monte Carlo, but the improvement is less impressive than for the efficiency. We can derive an enhanced variance reduction: $\mathrm{Var}_g\left(p_g^{\rm IS}\right) \le \kappa\mathrm{Var}_f\left(p_f^{\rm MC}\right) - \frac{1-\kappa}{N}p^2$ and thus $\sigma_g\left(p_g^{\rm IS}\right) \le \sqrt{\kappa}\,\sigma_f\left(p_f^{\rm MC}\right)$, which for $\kappa = 0.1$ means that here not an order is gained, but a factor $\sqrt{\kappa} \approx 0.316$.

We note that, if $g(x) \equiv 1$, as in Section 2, we have $\mathrm{Var}_g(V) = \frac{1}{pq}$, see (2). We remark that (4) is easily satisfied if f is a Gaussian distribution and g has a broader or shifted (Gaussian, or uniform) distribution, with enough density on A. In [2] one uses a 4σ shift for a Gaussian distribution; in [3] the shift is optimized. In [11] and in [4, 9] algorithms for an adaptively determined distribution g can be found.

4. Uncertainty Quantification

Uncertainty Quantification usually applies so-called Polynomial Chaos expansions of the random processes. The corresponding numerical approaches represent an alternative to do statistics, and are in many cases several orders faster than what is possible with Monte Carlo. Thus, statistics can be done efficiently, exploiting fast converging expansions, and with a sound mathematical background.

Around 2005 interest popped up in electronic engineering. In the Polynomial Chaos approach, one represents a solution by an expansion using orthogonal polynomials, where the polynomials involve the random parameters and the coefficients are time or space dependent. These coefficients have to be determined by some numerical technique, where mostly the two classes of Collocation and Galerkin methods are applied. On the one-hand, these techniques offer deterministic algorithms. On the other hand, they require either many systems to be solved (Collocation), or a large fully coupled system (Galerkin). The classical Hermite polynomials (associated with normal distributions) are worse in the tails; an expansion using Gauss-Legendre polynomials (associated with uniform distributions) already behaves better.

The software tool RODEO of Siemens AG seems to be the only industrial implementation of failure probability calculation that fits within the polynomial chaos framework [13]. The method can shift the (probability density) weighting function in the inner product to the area of interest (shifted Hermite chaos). One also can use a windowed Hermite chaos. The shift is tuned by some optimization procedure. The windowed Hermite chaos is the most accurate.

In [14] for a parameter $\gamma = \gamma_0 + \gamma_1 \xi$, where ξ is a beta random variable, one considers an expansion in Jacobi polynomials; more generaly, knowing the density of γ one can construct orthogonal polynomials.

A hybrid method to compute small failure probabilities has been introduced by [10], where the method achieves efficient numerical simulations for academic examples. Most likely, this technique has not been applied in European industrial companies yet.

5. Accurate Estimate of SRAM Yield

The threshold voltages V_t of the six transistors in an SRAM cell are the most important parameters causing variations of the characteristic quantities of an SRAM cell [5] like Static Noise Margin (SNM) and Read Current (I_{read}). In [5, 11] Importance Sampling (IS) was used to accurately and efficiently estimate low failure probabilities for SNM and I_{read}. $\text{SNM} = \min(\text{SNM}_h, \text{SNM}_l)$ is a measure for the read stability of the cell. SNM_h and SNM_l are identically Gaussian distributed. The min() function is a non-linear operation by which the distribution of SNM is no longer Gaussian. Figure 2-top, shows the cumulative distribution function (CDF) of the SNM, using 50k trials, both for regular MC (solid) and IS (dotted). Regular MC can only simulate down to $P_{\text{fail}} \leq 10^{-5}$. Statistical noise becomes apparent below $P_{\text{fail}} \leq 10^{-4}$. With IS (using a broad uniform distribution g), $P_{\text{fail}} \leq 10^{-10}$ is easily simulated (we checked this with more samples). The correspondence between regular MC and IS is very good down to $P_{\text{fail}} \leq 10^{-5}$. The Read Current I_{read} is a measure for the speed of the memory cell. It has a non-Gaussian distribution and the cumulative distribution is shown in Figure 2-bottom. Also here IS is essentially needed for sampling I_{read} appropriately.

Extrapolated MC assumes a Gaussian distribution based on estimated expectation and standard deviation (which only need a few number of samples). Figure 2-top clearly shows that using extrapolated MC (dashed) leads to overestimating the SNM at $P_{\text{fail}} = 10^{-10}$. Figure 2-bottom shows that extrapolated MC can result in serious underestimation of I_{read}. This can lead to over-design of the memory cell.

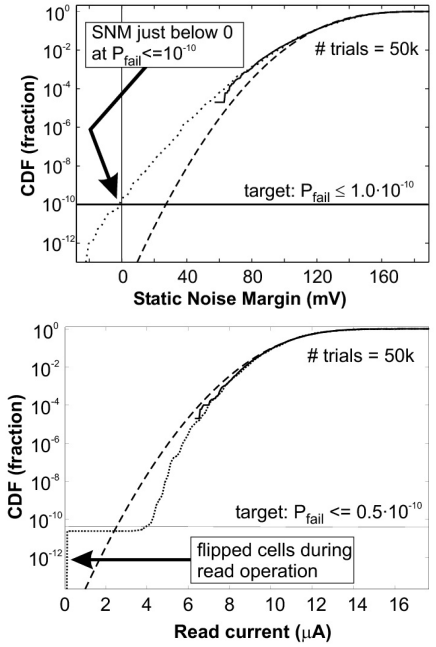

Fig. 2. SNM (top) and I_{read} (bottom) cumulative distribution function for extrapolated MC (dashed), regular MC (solid) and IS (dotted). Extrapolation assumes a normal distribution.

6. Optimization of SRAM Block

The block in Fig. 3 contains a Sense Amplifier (SA), a selector, and a number of SRAM cells. The selector chooses one "row" (block) of cells. Then the voltage difference is $\Delta V_{\text{cell}} = \Delta V_{\text{k}}$. A block B works if $\min_{\text{k}}(\Delta V_{\text{k}}) \geq \Delta V_{\text{SA}}$. With m blocks B and n cells per block we define Yield Loss by $YL = P(\#B \geq 1)$. Note that $P(\#B \geq 1) \leq mP(B)$, where the fail probability $P(B) = P_{\text{fail}}(B)$ of one block is (accurately) approximated by the lower bound $P(B) \approx \frac{YL}{m} = \frac{nYL}{N}$, in which $N = nm$. For $YL = 10^{-3}$, $m = 10^4$ blocks, $n = 1000$ we find $P(B) \leq 10^{-7}$.

For $X = \min_k(\Delta V_k)$, and $Y = \Delta V_{\text{SA}}$ we have

$$\begin{aligned} P(B) &= P(X < Y) = \int\int_{-\infty \leq x < y \leq \infty} f_{X,Y}(x,y)\,dx\,dy \\ &= \int_{-\infty}^{\infty} f_Y(y)\,F_X(y)\,dy. \end{aligned} \tag{6}$$

Thus we need the pdf $f_Y(y)$ and the cdf $F_X(y)$ (probability and cumulative distribution functions of Y and X). Note that

$$\begin{aligned} F_X(y) &= P(X < y) = P(\min_k \Delta V_k < y) \\ &= 1 - [1 - P(\Delta V_k < y)]^n \leq nP(\Delta V_k < y). \end{aligned} \tag{7}$$

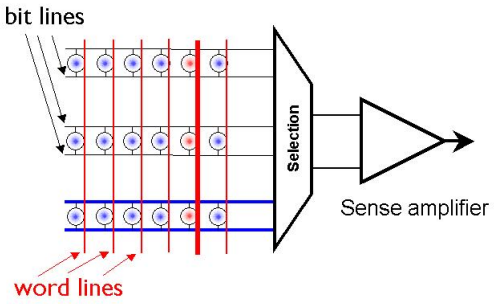

Fig. 3. Rows with blocks of SRAM cells together with a Selector and a Sense Amplifier.

For each simulation of the block we can determine the access times Δt_{cell} and Δt_{SA}. We come down to an optimization problem with a statistical constraint:

Minimize $\Delta t_{\text{cell}} + \Delta t_{\text{SA}}$ such that $P(B) \leq 10^{-7}$.

This has led to the following algorithm. We only give a sketch; for more details see [6].

- By **Importance Sampling** sample ΔV_k. Collect ΔV_k at the same Δt_{cell}.
- By **Monte Carlo** sample ΔV_{SA}. Collect ΔV_{SA} at the same Δt_{SA}.
- For given Δt_{cell}:
  - Estimate pdf $f_{\Delta V_k}$ and cdf $P(\Delta V_k < y)$.
  - From this calculate $F_X(y) = F_X(y; \Delta t_{\text{cell}})$, using the exact expression in (7). In our case we have $\frac{\partial F_X(y;\Delta t_{\text{cell}})}{\partial \Delta t_{\text{cell}}} \leq 0$.
- For given Δt_{SA}:
  - Estimate pdf of ΔV_{SA}: $f_Y(y)$.
- Calculate (numerical integration)
  - $P(B) = \int_{-\infty}^{\infty} f_Y(y) F_X(y) dy$.

Hence $P(B) = G(\Delta t_{\text{cell}}, \Delta t_{\text{SA}})$ for some function G. For given Δt_{SA} $G_1(\Delta t_{\text{cell}}; \Delta t_{\text{SA}}) = G(\Delta t_{\text{cell}}, \Delta t_{\text{SA}})$ is monotonically decreasing in Δt_{cell}. Hence we *Minimize* $G_1^{-1}(10^{-k}; \Delta t_{\text{SA}}) + \Delta t_{\text{SA}}$. The optimization with the statistical constraint on $P(B)$ led to a reduction of 6% of the access time of an already optimized SA while simultaneously reducing the silicon area [6].

7. Conclusions

We derived sharp lower and upper bounds for estimating accuracy of tail probabilities of quantities that have a non-Gaussian distribution. For Monte Carlo and for Importance Sampling (IS) this leads to a realistic number of samples that should be taken. IS was applied to efficiently estimate fail probabilities $P_{\text{fail}} \leq 10^{-10}$ of SRAM characteristics like Static Noise Margin and Read Current. We also applied IS to minimise the access time of an SRAM block while guaranteeing that the fail probability of one block is small enough. In our experiments we used a fixed distribution g in the parameter space. In [11] an algorithm with an adaptively determined distribution g can be found.

Acknowledgement: The 2nd and 5th author did part of the work within the project ARTEMOS (Ref. 270683-2), `http://www.artemos.eu/` (ENIAC Joint Undertaking).

References

[1] BUCKLEW, J.A., *Introduction to rare event simulation*. Springer, 2004.

[2] CHEN, G., SYLVESTER, D., BLAAUW, D., MUDGE, T., Yield-driven near-treshold SRAM design. *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, 18-11, 2010, p. 1590–1598.

[3] DATE, T, HAGIWARA, S., MASU, K., SATO, T., Robust importance sampling for efficient SRAM yield analysis. *Proc. ISQED'2010, 11th Int. Symp. on Quality Electronic Design*, 2010, p. 15–21.

[4] DONG, C., Li, X., Efficient SRAM failure rate prediction via Gibbs sampling. *Proc. Design Automation Conference (DAC) 2011*, p. 200–205 (12.3).

[5] DOORN, T.S., MATEN, E.J.W. TER, CROON, J.A., DI BUCCHIANICO, A., WITTICH, O., Importance Sampling Monte Carlo simulation for accurate estimation of SRAM yield. In: *Proc. IEEE ESSCIRC'08, 34th Eur. Solid-State Circuits Conf.*, Edinburgh, Scotland, 2008, p. 230–233.

[6] DOORN, T.S., CROON, J.A., MATEN, E.J.W. TER, DI BUCCHIANICO, A., A yield statistical centric design method for optimization of the SRAM active column. In: *Proc. IEEE ESSCIRC'09, 35th Eur. Solid-State Circuits Conf.*, Athens, Greece, 2009, p. 352–355.

[7] DE HAAN, L., FERREIRA, A., *Extreme Value Theory*. Springer, 2006.

[8] DEN HOLLANDER, F., *Large Deviations*. Fields Institute Monographs 14, The Fields Institute for Research in Math. Sc. and AMS, Providence, R.I., 2000.

[9] KATAYAMA, K., HAGIWARA, S., TSUTSUI, H., OCHI, H., SATO, T., Sequential importance sampling for low-probability and high-dimensional SRAM yield analysis. *Proc. IEEE ICCAD 2010*, p. 703–708.

[10] LI, J., LI, J., XIU, D., An efficient surrogate-based method for computing rare failure probability. *J. Comput. Phys.*, 230, 2010, p. 8683–8697.

[11] MATEN, E.J.W. TER, DOORN, T.S., CROON, J.A., BARGAGLI, A., DI BUCCHIANICO, A., WITTICH, O., Importance sampling for high speed statistical Monte-Carlo simulations – Designing very high yield SRAM for nanometer technologies with high variability. *Report TUE-CASA 2009-37*, TU Eindhoven, 2009, `http://www.win.tue.nl/analysis/reports/rana09-37.pdf`.

[12] MATEN, E.J.W. TER, WITTICH, O., DI BUCCHIANICO, A., DOORN, T.S., BEELEN, T.G.J., Importance sampling for determining SRAM yield and optimization with statistical constraint. To appear in: MICHIELSEN, B., POIRIER, J.-R. (Eds.), *Scientific Computing in Electrical Engineering SCEE 2010*, Series Mathematics in Industry Vol. 16, Springer, 2012, p. 39–48.

[13] PAFFRATH, M., WEVER, U., Adapted polynomial chaos expansion for failure detection. *J. of Comput. Physics*, Vol. 226, 2007, p. 263–281.

[14] SAFTA, C., SARGSYAN, K., DEBUSSCHERE, B., NAJM, H., Advanced methods for uncertainty quantification in tail regions of climate model predictions, Poster ID: NG31B-1324, Sandia National Laboratories, 2010.