Generative Model Based Adaptive Importance Sampling for Flux Calculations in Process TCAD

Alexander Scharinger*, Paul Manstetten†, Andreas Hössinger‡, and Josef Weinbub*
*Christian Doppler Laboratory for High Performance TCAD at the
†Institute for Microelectronics, TU Wien, Gußhausstraße 27-29/E360, 1040 Wien, Austria
‡Silvaco Europe Ltd., Compass Point, St Ives, Cambridge, PE27 5JL, United Kingdom
Email: scharinger@iue.tuwien.ac.at

***Abstract*—A key part of advanced three-dimensional feature-scale etching and deposition simulations is calculating the particle flux distributions. The most commonly applied flux calculation approach is top-down Monte Carlo which, however, introduces numerical noise. In principal, this noise can be reduced by increasing the number of simulated particles but doing so also increases the overall running time. For complex geometries, especially high aspect ratio structures, which are very prominent in state of the art three-dimensional electronic device designs, increasing the number of samples is not a viable approach: Only a very small subset of simulated particles contributes to reducing the noise in remote and obscured surface regions. We thus propose an adaptive importance sampling approach based on a generative model to more efficiently focus the sampling on those surface regions with high noise levels. We show that, for a constant number of simulated particles, our approach reduces the noise levels in the calculated flux by about 33% for a representative high aspect ratio test structure.**

I. Introduction

Three-dimensional feature-scale simulations of etching and deposition processes require an accurate simulation of the particle transport to obtain realistic particle flux distributions on the structure surface. A common approach to compute the particle flux distributions is to use a Monte Carlo simulation [1]. In particular, *top-down* Monte Carlo particle tracing is a popular and versatile approach because the integration of adsorption and reflection characteristics of any detail is straightforward and flexible [2]. A top-down Monte Carlo flux simulation consists of sampling (pseudo) random particles (origin and direction) from a given source distribution. Each particle carries a weight or energy value. The trajectories of the particles are traced through the simulation domain assuming ballistic transport. When a particle hits a material surface the local adsorption and reflection characteristics determine the effect of the particle on the local flux and its further treatment, e.g., re-emission of subsequent particles.

A major downside in conventional Monte Carlo particle tracing algorithms is the inherent numerical noise present in the results. This is particularly the case for (very) high aspect ratio (HAR) structures which are prominent in many semiconductor devices, e.g., NAND flash cells [3]. Even worse, the noise in the calculated flux distribution can be highly heterogeneous across the simulation domain [4], [5]. A straightforward but computationally very expensive way to deal with the noise is to excessively increase the number of simulated particles. However, this typically leads to massive oversampling in many parts of the surface while regions remote from the particle source and obscured from direct flux contributions still reveal high noise levels.

We present an adaptive Monte Carlo importance sampling algorithm applied to the source of the particles to reduce oversampling and to drastically decrease the noise levels in remote and obscured regions without increasing the total number of particles. In our approach, we start by distributing a limited number of particles in the conventional way to identify regions with high noise levels. We use this information to construct a Gaussian mixture model (GMM) [6], [7], which is an established generative model from the field of machine learning. The GMM's probability distribution is used to generate the remaining (major) share of particles, ultimately allowing to increase the accuracy in regions with high noise levels. In Section II we introduce the method and in Section III the method is applied and analyzed: Our approach reduces the overall noise in the calculated flux distribution on a representative HAR structure by 33% compared to the conventional Monte Carlo technique when maintaining the number of simulated particles.

II. Method

The proposed adaptive sampling algorithm for computing the flux distribution is illustrated in Fig. 1. It is inspired by the cross-entropy method for rare event simulation and combinatorial optimization [8].

The first stage consists of performing a conventional Monte Carlo flux calculation by sampling from the original uniform distribution at the source of the particles (Fig. 1a). This first stage yields for each surface element a flux estimate and a relative error of the estimate (Fig. 1b). The relative error is defined by $\sqrt{\mathrm{Var}\left(\hat{f}\right)}/\mathbb{E}\left(\hat{f}\right)$ ($\mathrm{Var}\left(\hat{f}\right)$ and $\mathbb{E}\left(\hat{f}\right)$ denote the variance and the expected value of the flux estimate $\hat{f}$, respectively) and quantifies the convergence of the Monte Carlo estimation [1].

The second stage uses a small number of particles to track which locations at the particle source hit surface elements with a flux estimate with high relative error (Fig. 1c). We call this stage *importance mapping* and it produces a set of relevant locations at the particle source.

Next, a GMM [6], [7] is fitted to the output of the importance mapping (Fig. 1d).

A GMM is a parametric probability density function characterized by the weighted sum of a finite number of Gaussian components: $p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ where K is the number of components, π_k the mixing weights, and $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the the mean vectors and covariance matrices of the Gaussian distributions $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, respectively. The mixing weights π_k, the mean vectors $\boldsymbol{\mu}_k$, and the covariance matrices $\boldsymbol{\Sigma}_k$ are GMM parameters. A GMM is a generative probabilistic model, that is, it can be used to generate new random data points reproducing its probability distribution. The particle source in the three-dimensional simulation domain is a plain surface, hence the Gaussian distributions are two-dimensional. Given the set of relevant samples from the importance mapping step, the parameters of the GMM are computed using maximum likelihood estimation and the expectation maximization algorithm [7]. The optimal value for K (the number of components in the GMM) is automatically selected via the the integrated completed likelihood (ICL) criterion [9]. Random sampling from a GMM requires only minor computational overhead compared to sampling from a uniform distribution using, e.g., the Box-Muller approach for Gaussian sampling [1].

The final stage of the algorithm is to use the probability distribution of the obtained GMM for Monte Carlo importance sampling [1]: Particles are generated according to the probability distribution of the GMM and each particle is weighted with the inverse sampling probability such that the Monte Carlo estimates remain unbiased (Fig. 1e), resulting in the final surface flux distribution and corresponding relative errors (Fig. 1f).

III. Results

We assess our algorithm with a representative HAR test structure shown in Fig. 2a. We compare conventional sampling using 32 M (million) particles with the proposed adaptive importance sampling algorithm using about 8 M particles for the first stage ($S_{\text{conv.}} \approx 8\,\text{M}$ in Fig. 1a), about 100 k (thousand) particles for the importance mapping ($S_{\text{imp.}} \approx 100\,\text{k}$ in Fig. 1c), and about 24 M particles for the importance sampling ($S_{\text{GMM}} \approx 24\,\text{M}$ in Fig. 1e), totaling again 32 M particles. For the importance mapping (Fig. 1c) we consider all surface elements with a relative error greater than 0.1 (i.e. 10%).

Fig. 3a shows the set of relevant samples obtained from the importance mapping step when running the proposed algorithm on the test structure. About 100 k samples for the importance mapping result in 634 relevant samples. It is apparent that the contributions from particles sampled right above the circular holes and the trench dominate the set of relevant samples.

From the set of relevant samples the fitting step (cf. Fig. 1d) results in a GMM with three components ($K = 3$). Fig. 3b visualizes the probability density function of this GMM (i.e. the superposition of three two-dimensional Gaussian distributions). This distribution is used for Monte Carlo importance sampling in the importance sampling step (cf. Fig. 1e). Although the geometry is axis-symmetric, the distribution of the GMM in Fig. 3b is not strictly symmetric. As the set of relevant samples is produced using a Monte Carlo approach these deviations from symmetry are expected.

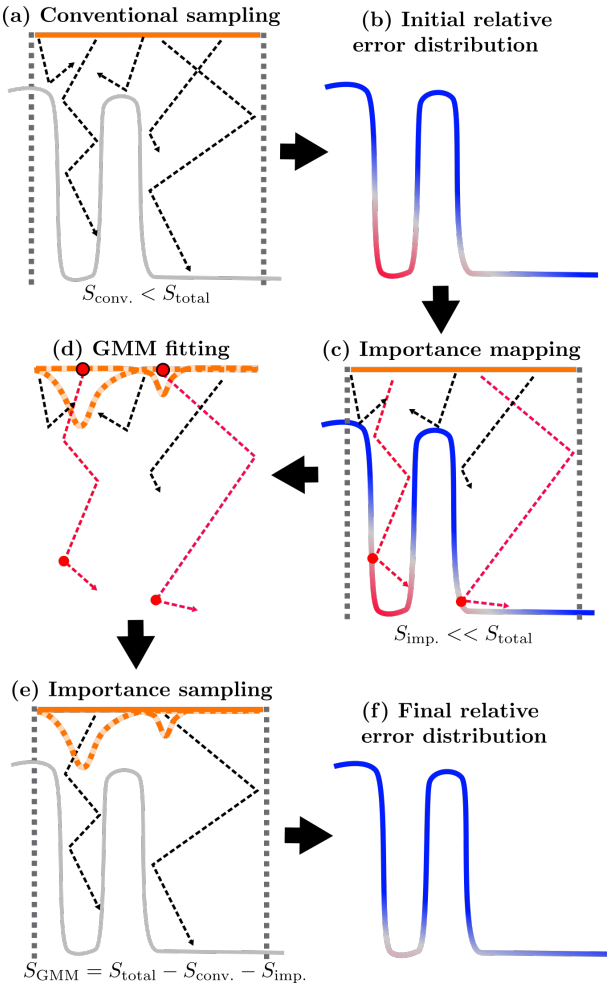
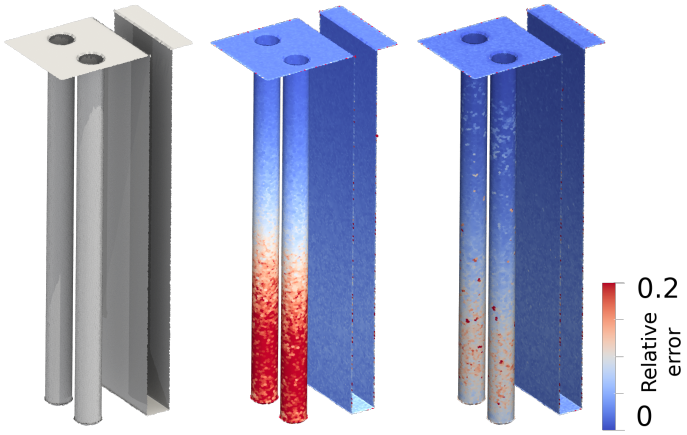


Fig. 1: Schematic sequence of the proposed adaptive importance sampling algorithm: Instead of a full conventional sampling of S_{total} particles, only a portion of particles $S_{\text{conv.}}$ is sampled conventionally to generate an initial error distribution. This initial distribution is used to identify important samples from a small subset $S_{\text{imp.}}$, which are used to fit a GMM. Finally, the majority of particles S_{GMM} is generated using importance sampling based on the distribution of the fitted GMM.

Fig. 2b shows the relative error of the flux estimates using conventional sampling. The bottom of the two circular holes in the test structure shows high levels of relative error of the calculated flux. In comparison to this, Fig. 2c shows the relative error using the proposed algorithm when applying adaptive importance sampling. The error distributions are visualized in Fig. 4 which compares the distributions of the relative errors using conventional sampling and the proposed adaptive importance sampling: One can clearly see that our approach significantly reduces the number of surface elements with high relative error.



(a) (b) (c)

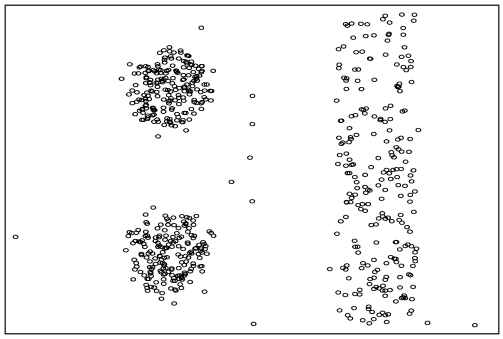
Fig. 2: HAR test structure consisting of about 40 k surface primitives: (a) The plain geometry, (b) the relative error of the surface flux distribution using 32 M conventionally sampled particles, and (c) the relative error of the surface flux distribution using 32 M particles sampled with the proposed adaptive importance sampling algorithm.

The noise in the calculated flux distribution is reduced by about 33% (from 0.3 to below 0.2 in Fig. 4) for surface regions with high noise in remote and obscured areas of the geometry.

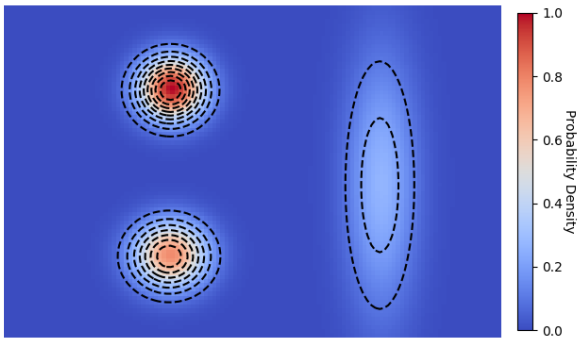
To provide a statistically unbiased sampling of the particle source distribution, the weights of the particles sampled from the GMM exhibit a broad range of weights (energies). A very small number of these particles carry a very high weight value. These particles actually have a negative effect on the computed flux and its relative error: When they hit the surface they increase the relative error of the flux estimate at the surface element, especially in regions with low particle flux, e.g., towards the bottom of HAR structures. This effect is apparent in Fig. 2c: A few surface elements in the otherwise continuous gradient exhibit greater relative errors than others. A straightforward way to deal with high-energy particles would be to introduce a particle splitting scheme which can be found, for example, in [1].

IV. Summary and Outlook

We propose an adaptive importance sampling approach based on a generative model to more effectively sample particles in flux calculations. We assess our algorithm on a representative HAR test structure and show that it reduces the noise by about 33% – without increasing the number of Monte Carlo samples. A promising idea to improve the applicability of this approach to a wide range of geometries is to apply the adaptive importance sampling steps iteratively in order to progressively improve the focus of the importance sampling until a desired noise level is obtained for all surface elements.



(a) Set of relevant samples.



(b) Probability density function of the GMM ($K = 3$).

Fig. 3: Top view onto the HAR test structure shown in Fig. 2: (a) Visualization of the set of relevant samples (origins on source plain) computed in the importance mapping step of the proposed algorithm (Fig. 1c), (b) Visualization of the probability density function of the GMM for the set of relevant samples shown in (a) computed in the fitting step of the proposed algorithm (Fig. 1d).

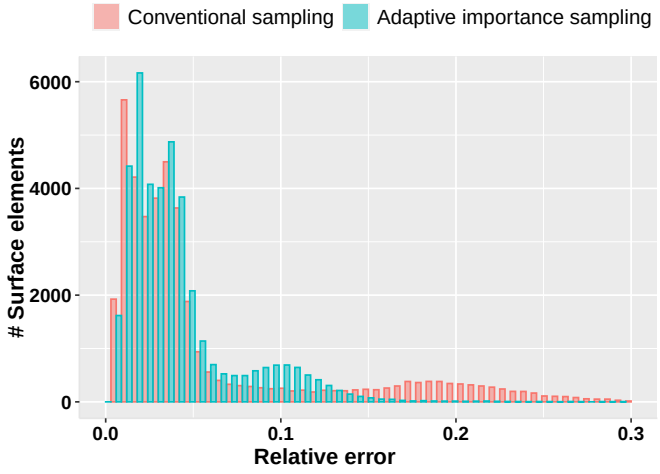


Fig. 4: Histogram of the relative errors of the flux estimates for the test structure with 32 M conventionally sampled particles as in Fig. 2b and 32 M particles with the proposed adaptive importance sampling algorithm as in Fig. 2c.

ACKNOWLEDGMENT

The financial support by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development, and the Christian Doppler Research Association is gratefully acknowledged.

REFERENCES

[1] R. Rubinstein and D. Kroese, *Simulation and the Monte Carlo Method*. Wiley & Sons, 2016.

[2] X. Klemenschits, S. Selberherr, and L. Filipovic, "Modeling of Gate Stack Patterning for Advanced Technology Nodes: A Review," *Micromachines*, vol. 9, no. 12, p. 631, 2018.

[3] P. Dimitrakis, *Charge-Trapping Non-Volatile Memories: Volume 1 – Basic and Advanced Devices*. Springer International Publishing, 2015.

[4] O. Ertl and S. Selberherr, "Three-Dimensional Level Set Based Bosch Process Simulations using Ray Tracing for Flux Calculation," *Microelectronic Engineering*, vol. 87, no. 1, pp. 20–29, 2010.

[5] P. Manstetten, L. Filipovic, A. Hössinger, J. Weinbub, and S. Selberherr, "Framework to Model Neutral Particle Flux in Convex High Aspect Ratio Structures Using One-Dimensional Radiosity," *Solid-State Electronics*, vol. 128, pp. 141–147, 2017.

[6] G. J. McLachlan and D. Peel, *Finite Mixture Models*. Wiley & Sons, 2004.

[7] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT press, 2012.

[8] P. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A Tutorial on the Cross-Entropy Method," *Annals of Operations Research*, vol. 134, no. 1, pp. 19–67, 2005.

[9] C. Biernacki, G. Celeux, and G. Govaert, "Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 7, pp. 719–725, 2000.