High-Sigma Performance Analysis using Multi-Objective Evolutionary Algorithms

***Abstract*—Semiconductor devices have rapidly improved in performance and function density over the past 25 years enabled by the continuous shrinking of technology feature sizes. Fabricating transistors that small, even with advanced processes, results in structural irregularities at the atomic scale, which affect device characteristics in a random manner. To simulate performance of circuits comprising a large number of devices using statistical models and ensuring low failure rates, performance outliers are required to be investigated. Standard Monte Carlo analysis will quickly become intractable because of the large number of circuit simulations required. Cases where the number of samples exceeds 10^6 are known as "high-sigma problems". This work proposes a high-sigma sampling methodology based on multi-objective optimisation using evolutionary algorithms. A D-type Flip Flop is presented as a case study and it is shown that higher sigma outliers can be reached using a similar number of SPICE evaluations as Monte Carlo analysis.**

I. Introduction

In order to accurately predict the performance distribution, let alone failures occurring with low probabilities, of circuits made of atomic-scale devices large numbers of time-consuming statistical SPICE simulations are required. In addition, realistic predictions with sufficient accuracy can only be made using specific, statistically enhanced device models [3], [10].

The most common state-of-the-art statistical method to sample the model parameter space is Monte Carlo (MC) analysis. However, if the circuit comprises of a large number of devices or model parameters (which is particularly the case for statistically enhanced device models) and very low failure rates or knowledge of performance outliers are required, MC analysis quickly becomes intractable because of the very large number of samples, i.e. circuit simulations, that are required to capture the tails of the performance distribution with sufficiently high probability. Cases where the number of samples required exceeds 10^6 are often called "high-sigma problems", reflecting the need to sample the tails of the distribution up to N sigma away from the median performance of a circuit ($N > 6$).

Due to the importance of efficient analysis of high-sigma problems, i.e. accelerating high-sigma sampling, it has been a research focus for several years and a number of methods for tackling high-sigma problems have been devised. Some of the most widely used approaches include principal component analysis, mixture importance sampling [4], [8] and minimum-norm importance sampling [5], which make the search space smaller through identifying the most critical parameters and components using sensitivity analysis and generate random samples that are then considered. Other approaches target the dimensionality of the search space more directly, e.g. in [2], [13] a series of searches are performed in lower-dimensional parameter spaces with subsequent application of a greedy search amongst the results. Another related method is Gibbs sampling [6], which is effectively MC analysis, but rather than generating samples randomly, Markov chains are used to generate sequences of samples which are correlated with nearby samples. Effectively this aims to approximate a joint probability distribution when sampling many parameters from a multivariate probability distribution is not easily possible. Alternative methods aiming to accelerate the mapping between parameter and performance space use surrogate models [7], [12] such as radial basis functions [11] or machine learning techniques [9] which build-up a model that can be executed fast to assist sampling large search spaces.

This work proposes a high-sigma sampling methodology based on multi-objective optimisation in conjunction with evolutionary (population-based) search algorithms [1]. The proposed method is in contrast to methods that explicitly tackle dimensionality reduction or optimised sampling from multivariate PDFs using formal statistical analysis, but it is related to surrogate-model-based approaches and complements mixture importance sampling techniques. In this work 25nm High-Performance Metal Gate Bulk MOSFET compact models from Gold Standard Simulations (GSS) (www.goldstandardsimulations.com) are used.

II. Methodology

A multi-objective evolutionary algorithm (MOEA) is used in this work for design optimisation, which is based on NSGA-II [1]. The distinctive feature of the proposed method is that an indirect encoding of the model parameters of each transistor is used, i.e. rather than randomly sampling from a probability distribution of a specific model parameter a *sampling strategy* for each device is optimised. In this case the strategies to chose from are simple: "-1" represents *chose a model card with this parameter value smaller than the current one*, "+1" represents *chose a model card with this parameter value larger than the current one* and "0" means *retain the current parameter value (model card)*. Device model cards are retrieved from a database of 10,000 PMOS and NMOS transistor model cards, respectively, which are generated using GSS's RandomSpice2 tool.

Hence, the optimisation algorithm's goal is to find model card selection strategies specific to every transistor in a circuit, which guide the search as quickly as possible to the tails of the performance distribution. This is achieved by calculating the fitness of a candidate circuit based on its performance measured in SPICE when traversing the model parameter space using the optimised sampling strategies. An example of this is shown in Figure 1. The use of a MOEA allows optimisation of a number of performance characteristics at the same time. The sampling strategies can be vectors of -1,0,1 rather than just scalars, taking multiple model parameters into account at the same time. In this initial work, however, the parameter space is kept one-dimensional for simplicity taking only V_{th} into account, although other parameters are updated as well keeping model cards consistent.

A population size of 100 is used, i.e. 100 netlists are generated per generation. If measuring a candidate circuit's performance after applying its model parameter sampling strategies yields a result that lies further out in the tail of the distribution it remains unchanged, otherwise a mutation operation is performed on the sampling strategies whereby a randomly chosen strategy is replaced with a random value of either -1, 0 or 1. Results shown are obtained after running the MOEA for 200 generations, which corresponds to 20,000 SPICE runs, as shown in Figures 1 and 2.

III. Case Study: D-Type Flip Flop

Delay and dynamic power consumption are considered as performance characteristics of a DFF in this case. We assume that the compact models used capture the effects of stochastic variability with sufficient accuracy to make realistic performance predictions of designs when fabricated. The GSS model generator processes SPICE netlists in such a way the netlist reflects one specific scenario

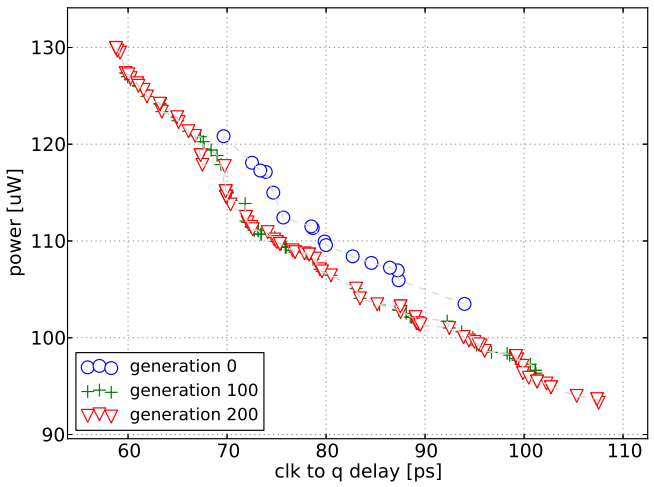

Fig. 1. Samples moving towards the tails of the performance distribution of the DFF. Generation 100 shows how the boundary has moved since generation 0. Generation 200 shows how the range has expanded since generation 100.

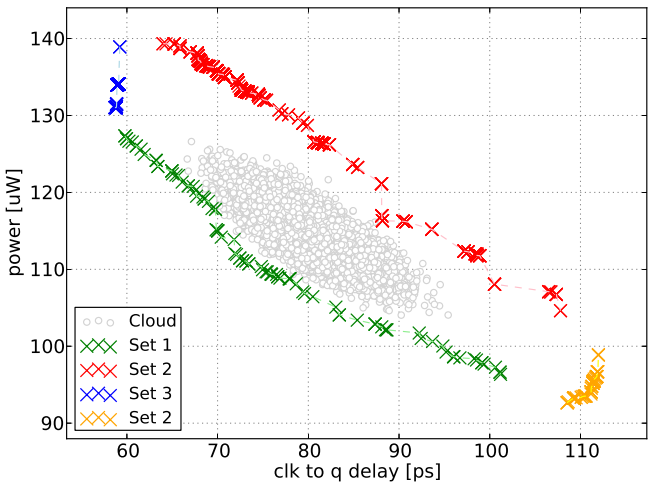

Fig. 2. *Cloud:* 20,000 VPIs of the DFF using MC and performance outliers when min. delay/power (*Set 1*), max. delay/power (*Set 2*), min. delay/max. power (*Set 3*) and max. delay/min. power (*Set 4*).

of intrinsic stochastic variability after processing. Hence, netlists generated can be regarded as possible physical implementations of a design, which allows inference of post-fabrication performance. Each unique randomised netlist of the DFF is therefore referred to as "virtual physical instance (VPI)". This process needs to be repeated many times in order to generate a large number (ideally thousands) of different netlists representing different points of a distribution of resulting design characteristics.

For comparison with the proposed high-sigma sampling method, 20,000 VPIs of the DFF are randomly generated measuring delay and dynamic power consumption using a SPICE testbench. The results are shown as scatter plot point cloud in Figure 2. The density, shape and spread of the point cloud illustrates the resulting variability distributions from the 20,000 simulation runs. Also shown in Figure 2 are the results of the high-sigma sampling performance of the MOGA-based approach described in Section II. In this case the four quadrants relating to min. delay/power, max. delay/power, min. delay/max. power and max. delay/min. power have been tackled in four independent runs amounting to a total of 40,000 SPICE runs.

IV. Conclusion

A multi-objective evolutionary optimisation technique has successfully been applied to the optimisation of device model sampling strategies accelerating high-sigma performance sampling of a DFF, a circuit with a significant number of transistors. The distance of samples from the mean could be doubled in all directions using only twice as many SPICE runs than MC. Using MC to get to the same sigma range would require up to 2 million SPICE runs.

Of course, we have only considered V_{th} in the optimisation process, which might make the problem somewhat simpler. However, the full parameter range of GSS model cards with correct model correlations was used and it is remarkable that considering V_{th} alone seems to yield significantly better results than standard MC analysis already.

The next steps will be to quantify our results, i.e. calculate the sigma achieved, and to increase the dimensionality of the parameter space during search. Based on our previous experience with MOGAs, we are confident that increasing the dimensionality of the performance space by at least 2-3 more metrics will still be feasible and sufficiently fast to evaluate, however, due to the complex relationships between the probability distributions of the model parameters this will be subject to further investigation.

References

[1] K. Deb and T. Goel (2001), Controlled Elitist Non-dominated Sorting Genetic Algorithms for Better Convergence. In: First Int. Conf. on Evolutionary Multi-Criterion Optimization, Springer, LNCS, no. 1993, pp. 67–81.

[2] M. Zhang, Z. Ye and Y. Wang (2014), Efficient high-sigma yield analysis for high dimensional problems. In: Design, Automation and Test in Europe Conference and Exhibition (DATE), pp. 1–6.

[3] A. Asenov (2007) Variability in the Next Generation CMOS Technologies and Impact on Design. In: Proceedings of the 1st International Conference of CMOS Variability.

[4] R. Kanj, R. Joshi and S. Nassif (2006), Mixture importance sampling and its application to the analysis of sram designs in the presence of rare failure events. In: ACM/IEEE Design Automation Conference, ACM, pp. 69-72.

[5] L. Dolecek, M. Qazi, D. Shah and A. Chandrakasan (2008), Breaking the simulation barrier: SRAM evaluation through norm minimization. In: IEEE/ACM Int. Conf. on Comp.-Aided Design, ACM, pp. 322-329.

[6] C. Dong and X. Li (2011), Efficient SRAM failure rate prediction via Gibbs sampling. In: ACM/IEEE Design Automation Conference, ACM, pp. 200-205.

[7] J. Yao, Z. Ye and Y. Wang (2013), Efficient importance sampling for high-sigma yield analysis with adaptive online surrogate modeling. In: Design, Automation Test in Europe Conference Exhibition, ACM, pp. 1291–1296.

[8] A. Singhee and R. Rutenbar (2007), Statistical blockade: A novel method for very fast Monte Carlo simulation of rare circuit events, and its application. In: Design, Automation Test in Europe Conference Exhibition, DAC, pp. 1–6.

[9] T. McConaghy (2011), High-dimensional statistical modeling and analysis of custom integrated circuits. In: IEEE Custom Integrated Circuits Conference (CICC), IEEE, pp. 1–8.

[10] K. Agarwal and S. Nassif (2008), The impact of random device variation on sram cell stability in sub-90-nm cmos technologies. In: IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 16, no. 1, IEEE, pp. 86–97.

[11] L. M. Surhone, M. T. Timpledon and S. F. Marseken (2010), Radial Basis Function Network. VDM

[12] K. Price, R. M. Storn and J. A. Lampinen (2005), Differential Evolution. A Practical Approach to Global Optimization. Springer, Berlin.

[13] M. Qazi, M. Tikekar, L. Dolecek and D. Shah (2010), Loop flattening & spherical sampling: Highly efficient model reduction techniques for SRAM yield analysis. In: Design, Automation Test in Europe Conference Exhibition (DATE), pp. 801–806.