Machine Learning for Microcontroller Performance Screening

Machine Learning for Microcontroller Performance Screening

Nicolò Bellarmino,
Dip. Automatica ed Informatica
Politecnico di Torino
Turin, Italy
nicolo.bellarmino@polito.it

***Abstract*—In safety-critical applications, microcontrollers must satisfy strict quality constraints and performances in terms of $F_{\max}$ (the maximum operating frequency). Traditional speed-binning techniques are not feasible to be applied to mass production, due to the high cost of the needed test equipment. Literature has proven that data extracted from on-chip ring oscillators (ROs) can model the $F_{\max}$ of integrated circuits by means of machine learning models able to predict the actual operating frequency of the devices. Those models, once trained, can be easily applied to the ROs data coming from every produced device with low effort and no need for high-cost equipment. This research aims to develop machine learning methodologies to be deployed in the MCU screening process, allowing for a more efficient and accurate $F_{\max}$ estimation, as well as improved speed binning. The effectiveness of this approach has been demonstrated on a real-world dataset of microcontroller data.**

***Index Terms*—Fmax, Speed Monitors, Ring Oscillators, Speed Binning, Machine Learning, Device Testing, Manufacturing, Semi-Supervised Learning, Deep Learning**

I. Introduction

Automotive and aerospace industries place a strong emphasis on the reliability of electronic devices, particularly in microcontrollers (MCUs) used in safety-critical components. To ensure that these devices perform as expected, MCU performance screening is used to identify any underperforming devices that do not meet the specifications outlined in their datasheets, specifically in terms of maximum operating frequency ($F_{\max}$). To determine the $F_{\max}$ of a device, it is tested under various worst-case conditions. Process variation during manufacturing impacts several parameters of integrated circuits, thus the performance of chips may vary across the production. Chips thus can be placed into different speed bins, depending on the performance (and chips with higher performance lead to more profit). To maximize profits, it is important to accurately and efficiently test the chips to place them in the correct bin, by performing tests at $F_{\max}$ speed, which can be divided into functional, structural (scan-based), and sensor-based tests [1]. Traditional functional methods of speed-binning involves running critical functional tests on the devices at increasing clock frequencies until a failure occurs. This permits measuring the $F_{\max}$. But this approach requires the use of high-end automated test equipment (ATE) to apply and analyze a large number of test patterns at high speed, resulting in high test overhead. Also, this approach consumes a large amount of memory on the tester and requires the use of costly ATE that can operate at the targeted frequency [1]. As chips designs become more complex and circuits faster the costs associated with functional testers are becoming prohibitive. Both the extended test time and the high requirement of ATEs increase the cost of traditional speed binning, making it unfeasible to be applied to large mass production. An alternative approach is to use machine learning (ML) regression models trained on data that can be correlated with the device's $F_{\max}$. In the literature, various methods for performance prediction have been suggested [2], [3]. The idea of using machine learning models to link structural and functional $F_{\max}$ was first introduced in [4]. Research has shown that on-chip ring oscillators, known as Speed Monitors (SMONs), can be used as features to predict $F_{\max}$ values from the execution of functional patterns on individual devices [5]. These ML regression models make it possible to relate the SMONs value with the continuous-value $F_{\max}$ of the devices. Using indirect measures to predict a circuit characteristic is called 'alternate test' in literature and has been widely studied for analog circuits [6], [7]. This approach leads to a mapping between indirect measurements and circuit specifications so that during production testing, device specifications can be predicted using only low-cost indirect measurements. ML is a promising approach in MCU performances, potentially deployable to mass-production, acquiring data from the SMONs of every produced device and predicting the operating frequency with high accuracy. However, the accuracy of supervised ML models depends on the quality and the quantity of labeled data. In this context, unlabeled data are relatively inexpensive to acquire. Instead, the acquiring $F_{\max}$ is costly in terms of time required, and thus the amount of available data is limited. The key difficulty in this context is obtaining a properly labeled training set (thousand of devices). This could require several months. The scarcity of labeled data requires carefully producing hand-crafted features and choosing simple models able to catch non-linear relationship in the $F_{\max}$ prediction. Also, noise in the data acquisition may affect the data quality. These are the main open problems. This research aim to develop ML models that take into account these critical issues, optimizing the labeling phase of the MCU sample and increasing the reliability of the overall procedure. The developed models are tested on several real case studies.

II. Contribution

The study in [5] firstly linked the data from 27-speed monitors to the functional $F_{\max}$ measurements on over 4,000 packaged devices taken from 26 corner-lot wafers, laying the foundation for a new area of research. A first step towards reducing the training set size for ML performance screening was moved in [8], by using Active Learning (AL) to choose the most informative samples for creating an ML model. AL techniques are deployable in contexts in which labeled data are hard to obtain, due to money or time cost. AL builds an optimized training set by selecting the most valuable sample, according to some metric, from a pool of unlabeled data. The selected sample will be labeled and inserted in the training set. This permitted halving the number of samples needed to build models, saving months in the overall procedure. In [9], the effectiveness of several outlier detection techniques was evaluated in identifying anomalous, noisy data, and outliers; as a result, the training set size was increased by recovering incomplete samples (samples with some missing values), which a classical ML framework would drop, obtaining higher quality data for feeding the successive ML models. This procedure permitted reducing the number of samples to be characterized (a third), without affecting the generalization error of the models. In [10], a novelty detection procedure based on wafer-level information derived by SMONs measurements was exploited, to estimate possible shifts in the data distribution. Monitoring the process permits test engineers to detect data (single wafers or entire lots) on which our model would have poor prediction performance, eventually labeling that data, and deciding if re-train or not the underlying prediction models. This procedure permits increasing the reliability of the prediction on mass production, and thus the robustness of the overall screening procedure. In general, Unsupervised or Semi-Supervised learning techniques can be deployed to take advantage of the ease of obtaining unlabeled data (possibly, for every produced device, and thus millions of samples). These techniques can extract relevant patterns, using these in the successive supervised performance prediction task. Deep Learning models, such as Auto-Encoders (never used in this context, by far) are appropriate to this goal since can be fed with the huge amount of available unlabeled data, and can be fine-tuned with labeled ones, thus becoming more effective and permitting to reduce the training set size required to train robust models. Noise detection techniques can also improve the quality of data, and thus contributes to fitting better models. All these methods are deployable to build more accurate models, increasing the yield of the screening process and thus the monetary gains, while reducing the testing time (that would require several months just to create the dataset, resulting in an infeasible application on large-scale volumes). But these techniques are quite general: they can be applied to many data-analysis other fields in which obtaining labeled data is hard (such as Biomedical data, Speech recognition, and Natural language processing), making this research topic extremely useful in a wide variety of contexts.

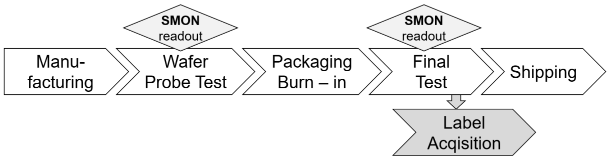

Fig. 1. Data collection steps through the manufacturing

III. Impact

The reasons why ML is a promising approach in microcontroller performance screening were presented above. It enables the use of models able to predict the operating frequency of the devices with high accuracy, and it can be potentially deployed to mass production. ML techniques have the potential to save significant test time and a huge amount of money (since requires no demanding ATE) with respect to traditional speed binning approaches [11]. However, the ML approach carries some uncertainty due to the probabilistic nature of the models, which depends on the data on which they are trained. But this uncertainty can be taken into account, by deploying guardband [5] to the output predicted frequency of the models, or by checking if the trained model is still appropriate to be applied to mass production [10]. With these approaches, ML models can achieve high prediction accuracy and quality (currently, about 1% of normalized mean absolute error, which is satisfactory for the proposed scope), and thus can be used in the traditional test flow of electronic devices, to support other existing testing procedures such as functional and structural tests, increasing the devices test accuracy.

References

[1] J. Zeng *et al.*, “On correlating structural tests with functional tests for speed binning of high performance design,” in *2004 International Conferce on Test*, 2004.

[2] K. von Arnim *et al.*, “An effective switching current methodology to predict the performance of complex digital circuits,” in *2007 IEEE International Electron Devices Meeting*, 2007.

[3] G. Sannena *et al.*, “Low overhead warning flip-flop based on charge sharing for timing slack monitoring,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2018.

[4] J. Chen *et al.*, “Data learning techniques and methodology for Fmax prediction,” in *2009 ITC*, 2009.

[5] R. Cantoro *et al.*, “Machine Learning based Performance Prediction of Microcontrollers using Speed Monitors,” in *2020 ITC*, 2020.

[6] H. Ayari *et al.*, “Making predictive analog/rf alternate test strategy independent of training set size,” in *2012 IEEE International Test Conference*, 2012.

[7] H.-G. Stratigopoulos *et al.*, “Error moderation in low-cost machine-learning-based analog/rf testing,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2008.

[8] N. Bellarmino *et al.*, “Exploiting active learning for microcontroller performance prediction,” in *2021 IEEE European Test Symposium (ETS)*, 2021.

[9] N. Bellarmino *et al.*, “Microcontroller Performance Screening: Optimizing the Characterization in the Presence of Anomalous and Noisy Data,” in *IEEE International Symposium on On-Line Testing and Robust System (IOLTS)*, 2022.

[10] F. Angione *et al.*, “Test, reliability and functional safety trends for automotive system-on-chip,” in *2022 IEEE European Test Symposium (ETS)*, 2022.

[11] S. Mu *et al.*, “Statistical Framework and Built-In Self-Speed-Binning System for Speed Binning Using On-Chip Ring Oscillators,” *IEEE VLSI*, 2016.