Accelerating Chip Design with Machine Learning: From Pre-Silicon to Post-Silicon

Cheng Zhuo\textsuperscript{1}, Bei Yu\textsuperscript{2}, Di Gao\textsuperscript{1},
\textsuperscript{1}College of ISEE, Zhejiang University, Hangzhou, China
\textsuperscript{2}Department of CSE, The Chinese University of Hong Kong, NT, Hong Kong
\textsuperscript{1}czhuo@zju.edu.cn

ABSTRACT

At sub-22nm regime, chip designs have to go through hundreds to thousands of steps and tasks before shipment. Many tasks are data and simulation intensive, thereby demanding significant amount of resources and time. Unlike conventional methodologies relying on experiences to manually handle data and extract models, recent advances in machine learning techniques enable the successful applications in various complex tasks to accelerate modern chip designs, ranging from pre-silicon verification to post-silicon validation and tuning. The goals are to reduce the amount of time and efforts to process and understand data through automatic and effective learning and enhancing from examples. In this paper we review and discuss several application cases of machine learning techniques, including pre-silicon hotspot detection through classification, post-silicon variation extraction and bug localization through inference, and post-silicon timing tuning through iterative learning and optimization, so as to leverage the potentials and inspire more future innovations.

1. INTRODUCTION

At sub-22nm regime, due to the growing process variation, increased design complexity and reduced design margin, chip design and manufacturing has become an overwhelmingly complex procedure that involves hundreds to thousands of steps \cite{1, 2, 3}. In order to ensure design robustness and performance target throughout entire procedure, designers and manufacturers have to conduct repeated tasks in pre-silicon verification, post-silicon validation and tuning \cite{4, 5, 6, 7, 8}. Many of such tasks demand processing, understanding and utilizing of a significant volume of data to verify chip functionality, validate process condition, and tune chip performance. Ignorance of any of the aforementioned data-intensive tasks may inevitably degrade chip design quality causing expensive over-design or impair manufacturing yield. On the other hand, conventional approach to such data (e.g., wafer process or lithography) and tasks is largely manual and ad-hoc, which is time and resource consuming and eventually delays time-to-market. Thus, it remains a critical field of research with exciting opportunities to deal with complex and data-intensive tasks for pre-and post-silicon chip design and optimization.

However, this is not a trivial work. Take post-silicon validation for example. Post-silicon validation typically begins with fabricated prototyped chips with data collected from each of them to understand the process condition, diagnose chip functionality and validate chip performance \cite{6, 9, 10, 11}. Even though testing can be conducted with high speed, due to the significant amount of data to be collected and extremely intensive tests to be conducted on each of the chips, the initial stages of validation are already slow. Moreover, once data is fed back, designers need to understand the data, associate that with underlying chip functionalities, root-cause and fix the problems if any. Since data and measurements can be easily impacted by process and environmental variations, it remains a challenge to derive effective models between chip functionalities and data. As a result, it is highly desired to have novel and efficient methods to deal with such data-intensive complex tasks at pre-and post-silicon stages to accelerate chip designs.

With recent advances in machine learning and its unique feature to extract structure from intensive data, many researchers \cite{3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14} proposed various promising approaches to handle complex problems ranging from pre-silicon verification to post-silicon validation and tuning. The goals are to reduce the amount of time and efforts to process and understand data through automatic and effective learning and enhancing from examples.

In the following, we will discuss the applications of various machine learning techniques in the aforementioned tasks to leverage their potential and inspire more future innovations. The discussion covers pre-silicon hotspot detection through classification, post-silicon variation extraction and bug localization through inference, and post-silicon timing tuning through iterative learning and optimization. The paper is organized as follows. Section 2 demonstrates the application of machine learning techniques in a pre-silicon verification problem of hotspot detection. Section 3 discusses the post-silicon process variation extraction and bug localization. Section 4 presents an example of post-silicon timing tuning through machine learning, followed by section 5 to conclude the paper.

2. PRE-SILICON HOTSPOT DETECTION

In physical design and verification stages, the hotspot detection problem is to locate hotspots on a given layout with fast turn-around-time. Conventional hotspot detection widely relies on full-chip lithography simulation, which can achieve very high accuracy but may suffer from being extremely computational expensive \cite{15, 16}.

To provide quick feedback to circuit designers, recently many fast and coarse-grained hotspot detectors are pro-
posed, which can be roughly classified into pattern matching based [17, 18, 19] and machine learning based (e.g., [20, 21, 22, 23]).

On one hand, pattern matching works on a pre-defined pattern set, thus it is less efficient on detecting unseen hotspots. Although fuzzy-pattern matching can be applied to dynamically tune appropriate fuzzy regions around known hotspots, it is still case sensitive and lacks the generalization ability to various detection environment. Some literature work combines both the pattern matching and machine learning methods. For instance, Wen et. al. [19] presents an idea of fuzzy matching integrated with some learning techniques, which can dynamically tune appropriate fuzzy regions around known hotspots in multi-dimensional space.

On the other hand, machine learning technique is to construct a regression model based on the training data set. Thanks to the good generalization ability, usually machine learning method can lead to higher prediction accuracy, with slightly longer training and prediction time. In addition, this method can recognize unknown hotspots through learning relationships between layout patterns and their printability. There are several works targeting at hotspot detection accuracy. Drmanac et. al. [20] propose a 2-D distance transform and histogram extraction method for mask layout images, which are used to construct the SVM-based hotspot detector. Ding et. al. [24] present a neural network judgment based detection flow, where 2-D hotspot patterns are directly used to train an ANN model. Ding et. al. [22] further propose a unified meta-classifier that enables several classifiers to work together. For each layout pattern, certain hotspot features are extracted and then fed into each base classifier, which calculates the prediction decision and generates a weight based on the weighting functions. In [23, 25], density-based layout features are fed into SVM engines for high performance hotspot detection. Reference [26] considers the hotspot detection in directed self-assembly (DSA) scenario, and enables robust learning with point correspondence and segment distance features. Zhang et. al. [27] develop smooth boosting based hotspot detection methodology, which can be further extended to online learning scenario.

It should be noted that feature extraction is always a key component in a high performance hotspot detector. References [28, 29, 30] perform feature extraction using principle component analysis (PCA), which significantly improves the performance of SVM-based hotspot detectors. In [31], thanks to the simplified layout features, the detector can avoid over-fitting issue, and is able to detect hotspots accurately with low false alarm. Zhang et. al. [27, 32] further study the feature optimization.

Very recently, some work develops deep learning-based framework targeting at high performance and large scale hotspot detection [7, 12, 13]. Because of the automatic feature learning technique and highly nonlinear neural networks, deep learning has shown improvement in term of accuracy and false alarm number especially on some large scale benchmark.

3. POST-SILICON VARIATION EXTRACTION AND BUG LOCALIZATION

With process shrinking below 22nm, chip manufacturing is subject to increasingly significant process variability [1, 33]. Thus, process variation models are considered critical not only for pre-silicon design, but also for post-silicon applications, such as characterization, validation, diagnosis and tuning, etc.. Due to its criticality, it has been a common practice in modern process to deploy hundreds to thousands of tiny test structures (e.g., ring oscillators and resistor arrays) within die or in the scribe line of the wafer for process monitoring and measurement [34, 35]. However, the overhead to measure all the test structures for all dies across all wafers is clearly too high [35, 36, 37]. By noting the statistical nature of process variation and the underlying process similarity from wafer to wafer, it remains an interesting yet challenging question how to reduce the testing cost by reusing measured information from earlier wafers to facilitate following wafer measurements.

To address this challenge, [9, 37, 38, 39] have proposed to simulate tens to hundreds of wafers to achieve a converged model and then use the model to reduce the number of measurements that were taken to monitor process on each wafer. However, in these approaches, the number of measurements and the choice of the measured devices were determined throughout the entire lot. In other words, they were not able to adapt to process change within the lot and from lot to lot. Instead of relying on fixed measurements, [10] proposed a Sparse Bayesian Learning based dynamic framework to monitor the process change and extract a variation model from small test structures. The extracted model considers various variation sources, including systematic variation, spatially-correlated variation and random variation. The proposed learning framework is capable to reduce the test cost by reusing the process model built from past wafer measurements and evolve the process model by taking partial measurements on the new wafer for validation and improvement.

Figure 1 provides an overview of the flow chart for the proposed framework. As can be seen from the figure, the framework starts with initial measurements conducted on the first (or first few) wafer(s) to train an initial process model that includes deterministic pattern (at wafer- and reticle levels) and non-deterministic variations. This model is achieved with high measurement density (i.e., measuring every possible test structure for this particular wafer). This measurement policy will later be changed with more wafers coming and model maturing. In other words, the number of measurements on each wafer is gradually decreased over time when the model fidelity increases. After training, for each wafer, a few reticles are selected to decide whether the process model is still valid or the process has gone through some unexpected change and needs adaption. After cross-validation, a small amount of tests is taken from untested reticles to estimate process variation parameters together with the trained models and further adapt the process model if necessary. Experimental results show that such measurement policy is able to significantly reduce the test cost while allowing the model to adapt and improve with the process changes. Similar as process variation extraction, bug localization in diagnosis, especially for analog and mixed signal circuits, is also confronted with a large number of possibilities, and almost impossible to execute all the paths and tests [6, 40, 41, 42]. Thus, it is not surprising the task is often manually handled by experienced designers with months of efforts [6, 41, 42].
Several machine learning and data mining techniques [6, 14, 43] have been proposed to extract potential invariants and variants in the model from the data. However, this still relies on a large amount of test data and suffers from the sensitivity to subtle process variations. Reference [11] proposes an approach to model the circuit using probabilistic graphical models and Bayesian inference with partially observed circuit responses. The approach is capable of modeling both circuit and environmental uncertainties and incorporating a small set of known data to the probabilistic graphical model to localize and rank the potential bugs. Although its efficiency may depend on the partial observations and induce false warnings, this still remains an effective approach to prune significant portion of the solution space, thereby making localization more tractable.

4. POST-SILICON TIMING TUNING

The growing process variations at sub-22nm have caused unaffordable timing margins during the design phase. These variations can be modeled as random variables, leading to statistical static timing analysis (SSTA) to analyze circuit yield [44]. In addition, new techniques ranging from circuit structure to special devices and mechanisms have been proposed to improve circuit performance or yield in view of these statistical characteristics. For example, the Razor method [45] increases the clock frequency until timing errors occur during operation.

Another way to alleviate the effect of process variations is to deploy post-silicon tuning devices into the circuit. After manufacturing, chips are adapted according to the final effect of process variations. Since this adaptation happens to each chip individually, chips with timing failures due to process variations have an additional chance to be rescued by a customized tuning scheme.

Post-silicon tuning can be applied in different domains. For example, a tunable delay buffer can be used to adjust the arrival time of a clock signal to the corresponding flip-flop. The structure of the delay buffer from [46] is shown in Figure 2, where the configuration bits in the three registers can be set through the scan chain to change the delay of the buffer. In high-performance designs, such buffers are inserted during the design phase. After manufacturing, more timing slack can be assigned to critical paths, which might be different in different chips, by tuning the delays of the buffers to shift clock edges toward the stages with smaller delays, so that the yield of the circuit can be improved [47, 48].

Since the size of a tuning buffer is quite large, it is not profitable to insert a tuning buffer for each flip-flop in the circuit. Instead, only the flip-flops related to the paths affecting the circuit performance/yield should be considered as candidates for buffer insertion. The straightforward technique to insert clock tuning buffers to critical paths, however, does not produce a good result, because tuning the clock edge at a flip-flop affects all the paths incoming to or outgoing from it. Considering the fact that all the gate/path delays should be considered as statistical when determining the buffer locations during the design phase, this problem can be formulated as a learning problem illustrated in Figure 3.

To determine the optimal locations of tuning buffers, the learning algorithm takes the circuit structure, the variation information from manufacturing process, supply voltage and operating temperature (PVT), as well as the results of physical design and parasitic extraction as input. From these data, the learning algorithm can construct the statistical characteristics of path delays in the circuit. Afterwards, these statistical delays are discretized to represent different manufacturing cases. For each sample, the learning algorithm evaluates potential buffer locations. Considering the probabilities of these samples, the learning algorithm gradually approaches the optimal combination of buffer locations while reducing buffer area and potential test cost.
In the research field of electronic design automation, machine learning techniques have been applied extensively, e.g., for test path selection under process variations [49], for placement of sensors to manage noise dynamically [50], and for yield calculation of analog-/mixed-signal integrated circuits [51]. Specifically for allocating post-silicon tuning buffers considering process variations, the method [52] first samples process variations. For each sample, a limited number of buffer locations are identified. The buffer locations that appear most often in the results are selected as the result. This method is extended in [8] to process multiple samples simultaneously, so that the relation between discretized samples is also incorporated. The execution efficiency of this learning process is improved by using a low-discrepancy sample sequences (Sobol sequence). After manufacturing, the inserted buffers can also be used to reduce testing time for clock tuning, as explained in [53]. These methods are, however, still evolved from the traditional Monte Carlo method, and a direct combination of machine learning algorithms with the sampling-learning concept can potentially improve the existing results significantly further.

5. CONCLUSIONS

The paper discusses a few novel applications of machine learning techniques in chip design and manufacturing. With a growing number of data intensive tasks, it is highly desired to fully unlock the potential of machine learning in classification, inference and optimization. Its successful applications may mitigate designers' workload to extract the key information from massive data set, eventually accelerating chip design procedure. However, there remain many challenges and future works in this field. For example, in most prior research, it is often not well discussed how to train and validate the model, which is apparently crucial in practice. Moreover, it remains an open question to apply machine learning to tasks with both data and simulation intensity. One example of such tasks is power delivery simulation and optimization, which has a huge design space and demands many rounds of expensive simulations from early design stage to chip sign off [50, 54, 55]. The simulation itself is typically data dependent and sensitive to any error introduced in the solution space. It is interesting and crucial to develop data-independent machine learning algorithms by taking advantage of circuit structure to accelerate the simulation and optimization procedure. The impact of such a solution may significantly change the traditional back-end design methodologies and tools.

6. REFERENCES


