Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

PROCESSING TIME AND MACHINE AVAILABILITY PREDICTION IN SEMICONDUCTOR MANUFACTURING USING NEURAL NETWORKS

Taki E. Korabi
Gerard Goossen
Abhinav Kaushik
Tijmen Tieleman
Jasper Van Heugten
Jeroen Bédorf

Mindsai, Inc.
101 Cooper St
Santa Cruz, California, USA

Shiladitya Chakravorty
Detlef Pabst
John Thomas

Globalfoundries, Inc.
400 Stonebreak Road Ext.
Malta, New York, USA

ABSTRACT

In partnership with GlobalFoundries we have significantly advanced Processing Time (PT) and machine availability prediction in fabrication plants, utilizing an attention based neural network. This model is integrated into an automated Machine Learning Operations (MLOps) pipeline consisting of data collection, preprocessing, training and deployment. The data is augmented with features such as chamber usage and process sequences. Compared to the current model, which calculates average processing times over a predefined context, our approach has reduced the Mean Absolute Error (MAE) of PT predictions by 43% to 80% across the crucial areas: Etch, Diffusion, and Deposition. The model also produces high quality predictions for the remaining tools. The model is in the process of being implemented in the fabrication process (FAB) to improve scheduling, dispatching, and improve crucial Key Performance Indicators (KPIs) such as cycle time and throughput.

1. INTRODUCTION

Semiconductor manufacturing, characterized by its complex, multi-stage processes, demands precise timings at each stage to optimize throughput and overall efficiency. A key metric for the fabrication process is the Processing Time; the duration a tool requires to process a job (batch, lot, wafer). Accurate PT predictions play a vital role in the overall manufacturing process, influencing task scheduling and dispatching, which in turn impact key FAB KPIs like cycle time and throughput. For scheduling and dispatching tasks, being able to predict the next available machine in a dispatching tool family is of great importance.

For processing time and machine availability prediction, GlobalFoundries currently uses a sophisticated prediction method that serves as a baseline. The method includes average processing time calculation based on various contextual features. These features include, but are not limited to, the tool, recipe, technology in use and product type. Despite its practicality, this method falls short of accounting for the dynamic elements of the production process, such as the usage of specific chambers and process sequencing. In this work we present a method that overcomes these limitations by the usage of machine learning. The solution uses the attention mechanism, well known for its use in Large Language Models (LLM), to overcome the inherent complexities and dynamic nature of semiconductor manufacturing.

An attention-based model excels in handling high-dimensional, non-linear data, and sequential patterns, characteristics intrinsic to semiconductor FAB operations. In addition to the existing set of features two new features are incorporated; the specific chambers used in the process and the sequence of operations on a tool. Experimentation indicates that those factors significantly influence the accuracy of PT predictions. This extended abstract provides an overview of the data processing pipeline, a brief description of the model, and a summary of the results.

2 DATA PROCESSING

The data processing pipeline prepares the neural network inputs. The process begins with data cleaning, removing columns of singular or non-inference values. The model's depth is enhanced by integrating features from previous lots, thus expanding the scope for processing time and machine availability prediction. An embedding technique is utilized to handle complex categorical features. The clean and structured dataset is divided into training, validation, and test sets using a time-based split, preserving the temporal integrity essential for real-time predictions using a deep learning model.

3 MODEL DEVELOPMENT

The application of a neural network model using attention to capture the sequential information, forms the crux of the presented methodology, designed to capture the complexities and temporal dependencies of the fabrication processes. Attention mechanisms are used to enhance the model, assigning varying weights to sequence parts, thereby focusing on the most relevant inputs during prediction, which yields the model understanding of data relationships.

The model incorporates sequences from previous lots to discern patterns such as machine drift and the impact of previous processes. These are crucial for accurate prediction of the processing time and machine availability. Model training is performed using an automatic hyperparameter optimization process which mitigates overfitting, enhances its generalization capabilities, and ensures the model effectively captures data relationships for precise predictions.

4 RESULTS SUMMARY

Table 1 summarizes the improvement observed between the new deep learning-based model and the existing solution.

Table 1: Comparison results between the two models.

Area	Family	Current model MAE (sec)	New model MAE (sec)	MAE Improvement
Etch	ETX	933.9	310.5	66.7%
Deposition	CVD	228.6	136.3	43.1%
	MDX	373.6	183.6	50.8%
Diffusion	EPI	3421.4	659.4	80.7%
	RTA	521.1	186.8	64.1%