Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

HYBRID SIMULATION AND LEARNING FRAMEWORK FOR WIP PREDICTION IN SEMICONDUCTOR FABS

Taki Eddine Korabi[1], Gerard Goossen[1], Abhinav Kaushik[1], Jasper van Heugten[1],
Jeroen Bédorf[1], and Murali Krishna[1]

[1]minds.ai, Inc., Santa Cruz, CA, USA

ABSTRACT

This paper presents a hybrid framework that combines discrete-event simulation (DES) with neural networks to forecast Work-In-Progress (WIP) in semiconductor fabs. The model integrates three learned components: a dispatching model, an inter-start time predictor, and a processing time estimator. These models drive a lightweight simulation engine that accurately predicts WIP across various aggregation levels.

1 INTRODUCTION

Forecasting WIP levels is critical in semiconductor manufacturing, affecting throughput, cycle time, and delivery reliability. Existing approaches are either simplified analytical models that fail to capture fab complexity or detailed DES that are costly to develop and maintain. Although machine learning offers new possibilities, it often lacks the structure and interpretability required for deployment. This paper presents the minds.ai Maestro Fab Model, a hybrid framework that embeds neural network models within a DES engine to predict WIP evolution across different aggregation levels (e.g., tool groups, layers, recipes). The system is lightweight, modular, and deployable in fabs, enabling what-if analysis, proactive decision-making, and integration with optimization tools to enhance operational efficiency and maximize fab throughput.

2 MAIN CONTRIBUTION: THE FAB MODEL

The Fab Model is a hybrid approach that integrates machine learning with discrete-event simulation. At its core, the framework combines three neural models, each trained on historical fab trace data:

- **Dispatching model:** Selects the next lot to start from the current queue, capturing the dispatch logic of the fab and tool selection rules.
- **Inter-start time model:** Predicts the time interval between two consecutive lots starting at a given tool (or aggregation level). This model defines when the next lot will begin processing, independently of the specific lot identity.
- **Processing time model:** Estimates how long the selected lot will remain in process before completing the current step (Korabi et al. 2025).

Together, these models drive a discrete event simulation that tracks lot progression through the fab. The Fab Model supports any desired tool aggregation level—from individual tools to tool groups or fab-wide areas—making it adaptable to practical use cases. Aggregation units are linked dynamically via lot flow data, enabling accurate simulation of reentrant and multi-step flows. The model uses only standard fab data: operation history, product flows, tool states, events, and dispatch logs. No physical layout or detailed tool models are needed, making it portable and low-maintenance. Its data-driven design allows fast retraining as fab conditions evolve. The Fab Model scales to complex fabs, with a modular architecture enabling independent training and updating of components. It offers better maintainability and speed than full DES and greater interpretability than black-box ML, though chaining models may lead to error propagation and require retraining under major fab changes.

3 MAIN RESULTS

For demonstration purposes, we evaluate the Fab Model on synthetic data designed to reflect realistic fab behavior and operational complexity. The top plot in Figure 1 highlights the model's high accuracy in forecasting wafer arrivals across selected manufacturing steps. It effectively captures work-in-progress (WIP) fluctuations and maintains predictive reliability across all steps. The bottom plot illustrates predictions at the lot level—the granularity at which the Fab Model operates—where arrival times at each step are accurately estimated. This per-lot accuracy contributes to the strong aggregated performance observed in the top plot. This level shows the full line, but other aggregations—like area or tool—are also possible.

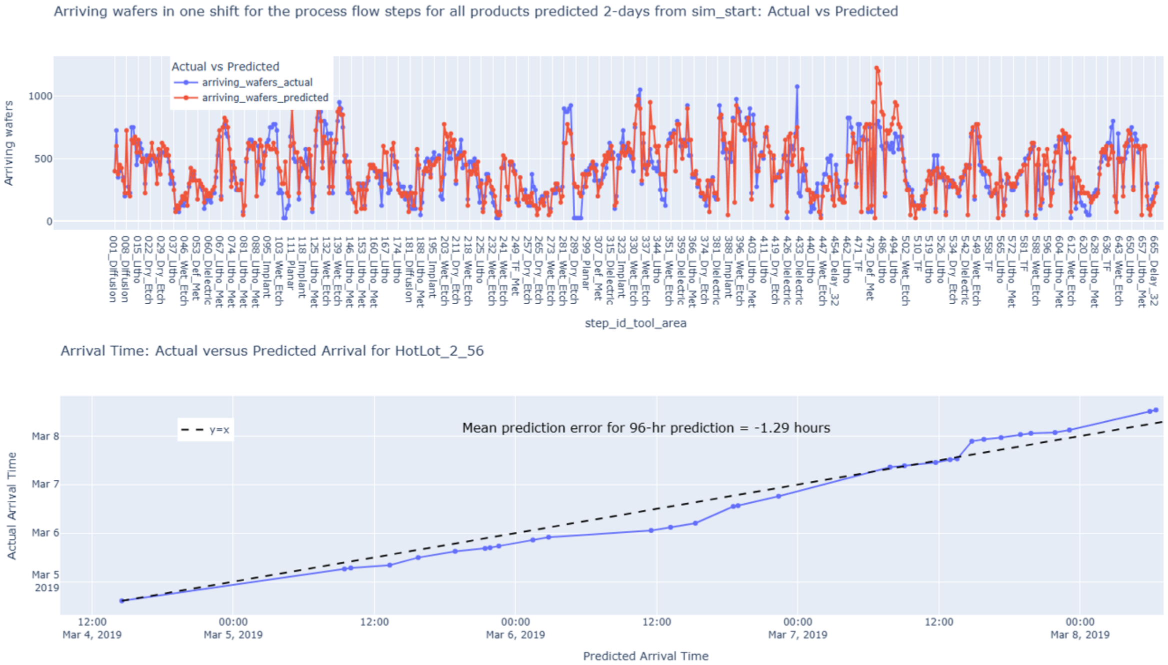

Figure 1: Fab Model prediction results. The top plot illustrates the predicted versus actual number of wafer arrivals at a selected subset of manufacturing steps, using a two-day lookahead horizon. The bottom plot presents a lot-level view, comparing the predicted and actual arrival times of the lot at various steps along its manufacturing route. The mean prediction error during the entire 4 days horizon is -1.3 *hours*.

CONCLUSION AND FUTURE WORK

The minds.ai Maestro Fab Model combines the flexibility of machine learning with the logic of discrete-event simulation to enable scalable and accurate WIP forecasting. This hybrid approach supports fast and proactive decision-making in complex manufacturing environments. The next steps include deployment in production facilities and integration with production planning systems.

REFERENCES

Korabi, T., G. Goossen, A. Kaushik, J. van Heugten, J. Bédorf, S. Chakravorty, *et al*. 2025. "General Framework for Processing Time Prediction and Machine Availability for all Fab Equipment". In *2025 36th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 1–6. IEEE.