Reinforcement Learning in Semiconductor Manufacturing: A Literature Review

Matthew S. Jones
Department of Electrical Engineering & Computer Science
Oregon State University
Corvallis, OR 97330
jonesm25@oregonstate.edu
www.mattjones.ai

March 16, 2025

Abstract

This literature review examines the use of reinforcement learning (RL) in a semiconductor manufacturing facility (fab). Key findings indicate an opportunity to leverage RL models to improve scheduling and process control. Opportunities include optimizing production paths to maximize throughput given equipment setup and maintenance schedules, and adjusting machine control parameters to react to disturbances in the environment. As compute capacity (for model training) expands, more complex RL models will be viable options for continued performance improvement and more precise control.

1 Introduction

Award-winning science fiction writer Sir Arthur C. Clarke once quipped, "Any sufficiently advanced technology is indistinguishable from magic." [Cla68] Nowhere in modern life is this more true than in semiconductor manufacturing. Combining sand and metal with chemicals to make a functioning integrated circuit containing billions of transistors is the most complicated manufacturing process in the history of mankind. The challenges that must be overcome to do this profitably at scale are as broad as they are technical. Newly cut wafers are polished, materials are added, microscopic patterns are created, different materials are added according to the pattern, the wafer is cleaned, then the cycle repeats on the same tools but with different materials and patterns (known as "re-entry") until processing is complete. The die are then cut and packaged before being shipped to customers. This literature review covers the latest RL-related research in semiconductor manufacturing by reviewed five recently published papers, primarily covering the biggest challenges of the past decade: efficient scheduling and advanced process control.

1.1 Search Parameters & Results

Three databases were searched for papers: *arχiv* at Cornell University, IEEExplore, and the ACM Digital Library (ACM-DL). The search parameters were "Reinforcement Learning" and "Semiconductor" and "Manufacturing," "Fab," or "Foundry" in the paper title, and published between 2017 and 2024. The five papers included in this review were primarily chosen for their diversity in publication venue. For example, there were many papers published in various years of the Proceedings of the Winter Simulation Conference, but only one was included in this review. The latest Machine Learning-focused conferences, NeurIPS and ICML, did *not* include research on semiconductor manufacturing.

1.2 Paper Organization

This paper is organized as follows: The next five sections (2 through 6) provide subsections for the problem addressed, a summary of the solution to the problem (which includes experiment setup and results), strengths and weaknesses, then take-aways and limitations, for each of the papers reviewed. The papers are ordered by publication year, then title. Section 7 concludes the paper by consolidating the key take-aways and proposing future research topics to further the use of RL in semiconductor manufacturing.

2 Robust Scheduling

Title: *A Reinforcement Learning Approach to Robust Scheduling of Semiconductor Manufacturing Facilities* [I B+20]; **Authors:** In-Beom Park, Jaeseok Huh, Joongkyun Kim, & Jonghun Park; **Publisher:** IEEE; **Publication:** Transactions on Automation Science and Engineering; **Year:** 2020.

2.1 Problem Addressed

Failure to efficiently schedule lots' movement through the manufacturing line can lead to poor machine utilization [Kov+23], equipment contamination and increased scrap [Cho+21], and lower overall factory yield. While the precise impact of poor scheduling and equipment setup timing cannot be quantified meaningfully because of the extreme variation between individual factories and product mixes, any improvement is valuable. As manufacturing of advanced semiconductor products increasingly includes multi-chip products (MCPs), this scheduling complexity has been extended to the wire bonding and packaging stages of the assembly process, which may vary in the number of machines they employ.

2.2 Solution Summary

The problem is defined as an NP-hard flexible job-shop (FJSP) that has been formulated as a Markov decision process (MDP) and implemented using a Q-

learning-based approach. The novel approach presented in the paper is to use a setup change scheduling method in which each RL agent determines setup decisions in a decentralized manner and learns a centralized policy. This is done by sharing a neural network among agents to deal with changes in the number of machines. The state and action representations of the MDP are designed to accommodate the variabilities in the production requirements and initial setup status. The authors then define the formula to calculate R (the sum of all rewards), which is maximized when C_{max} (the time of completion) is minimized.

The training phase employs a fully-connected neural network (NN) architecture for the Q-network. Specifically, the input and output are the state of a machine and the predicted action values of the individual actions, respectively. In the testing phase, the trained network is used to solve a test scheduling problem even when the production requirements, the number of machines, and the initial setup status of the problem differ from that of the training problems.

The experimental setup included 15 data sets that varied by the number of machines, the number of jobs, and the number of operations. Each data set has 30 scheduling problems whose production requirements were perturbed by 10% and the initial setup was randomly generated. The results were compared to a previously published genetic algorithm, which it outperformed by 1-32% across all data sets.

2.3 Strengths & Weaknesses

The primary strength and contribution of this solution is in the novel approach that provides flexibility and generalization. Independence from the number of machines provides an opportunity to simulate the impact of increased capital investment by predicting returns in the form of overall throughput. A lack of reliance on deterministic processing times provides additional flexibility. The primary weakness is the slow training time, which led the authors to choose more simplistic datasets than one would find in real-world applications.

2.4 Take-Aways & Limitations

This solution could be expanded to cover more than just wire-bonding and packaging, making it an effective solution for wafer processing as well. However, the solution did not appear to address significant machine downtime for maintenance, which is a reality that must be confronted by any effective scheduling algorithm. The challenge of slow training time will have to be addressed if this solution is to be viable for a real fab.

3 Process Control

Title: *Reinforcement Learning for Process Control with Application in Semiconductor Manufacturing* [LDJ21]; **Authors:** Yanrong Li, Juan Du, & Wei Jiang;

Publisher: IISE; **Publication:** Institute of Industrial and Systems Engineers Transactions; **Year:** 2021.

3.1 Problem Addressed

Process control technology is necessary for reducing the variation in manufacturing processes to improve quality and productivity of the final products. This variation can be caused by many different factors, including environmental (the altitude of and temperature inside the factory), tool-based (the age, maintenance status, or non-uniformity of the machine), or chemical (the quality and consistency of the consumable chemicals). Making these adjustments manually takes time and expertise, but still may not result in ideal run-to-run adjustments, leading to product quality issues or unnecessary (and expensive) scrap and waste.

3.2 Solution Summary

The stated problem is to prove that RL-based controllers have advantages over existing process controllers because of their ability to deal with complicated nonlinear processes. The contributions provided in this paper are: (1) a proposal of RL-based controllers, which can be applied in different situations, (2) two computational algorithms (one with domain knowledge, one without) for RL-based controllers, and (3) theoretical properties for RL-based controllers given the assumption of widely accepted linear process models:

1. The process model is defined as $y_t - f_t(y_{t-1}, u_t, d_t)$ for $t \in \boldsymbol{T}$, where d_t is the process disturbance of the system at time t, and u_t is the control action at period t. In each period there is an unavoidable disturbance d_t which can influence the system output, and the control action u_t aims to compensate for the effect of d_t.

2. When domain knowledge is present, statistical inference methods, such as maximum likelihood estimation can be used and the optimal control action can be determined. When domain knowledge is *not* present, a controller with policy gradient search (PGS), which estimates the distribution of input-output relationships from historical output data, is used instead.

3. A traditional controller requires knowledge of all parameters in the process model to begin, while an RL-based controller only requires the initial values. This gives the RL-based controller the ability to react to linear drift, which creates outliers using a traditional controller.

The comparison between traditional and RL-based controllers were done with two simulation cases: The Wiener process model and the Gamma process model, which represent two well-known stochastic processes. The simulations are done over 30 replications and the results presented are based on the mean and standard deviation of the mean squared error (MSE). The RL-based controller was shown to reduce process costs by more than 95%.

3.3 Strengths & Weaknesses

The primary strength of this solution is to clearly show the advantage of RL-based process control, when compared to traditional controllers. The paper includes a lot of formulae and pseudo-code to support the RL-based models and makes a very compelling case. Applying their solution to applications in chemical mechanical planarization (CMP) and deep reactive ion etching strengthens it. The primary weakness is that comparing an RL-based method to a traditional method is like comparing apples to oranges. The superiority of an RL-based solution was never in doubt.

3.4 Take-Aways & Limitations

This paper is a great primer on the theoretical basis for RL-based process control, providing theorems and proofs to support the claims. Limitations are the simplicity of the application of their solutions. More in-depth application may not have strengthened the case for RL-based controllers much more, but it would have more clearly shown their use in semiconductor manufacturing. An open challenge for this solution is to show performance advantages of their model over other RL models.

4 Run-to-Run Control

Title: *Distributional Reinforcement Learning for Run-to-Run Control in Semiconductor Manufacturing Processes* [MP23]; **Authors:** Zhu Ma & Tainhong Pan; **Publisher:** Springer; **Publication:** Neural Computing and Applications; **Year:** 2023.

4.1 Problem Addressed

Chemical mechanical polishing (also CMP since *polishing* and *planarization* are interchangeable) is a complicated process that involves the use of different types of slurries, multiple consumables, and varying downward pressures and relatively velocities (between the wafer and polisher) to remove non-uniformity from a semiconductor wafer's surface. In order to ensure a consistent result, run-to-run control (that is, changes at the end of each run in preparation for the next run) has been widely used and supported through deep reinforcement learning (DRL) techniques that rely on established models based on shift and drift disturbances. These models, however, do not take into consideration the nonlinear and time-varying nature of the process, so their accuracy is limited. Note: The problem addressed here revisits the problem addressed in **Process Control** [LDJ21] above.

4.2 Solution Summary

The solution presented in this paper is a new distributional DRL controller, quantile option structure deep deterministic policy gradient (QUOTA-DDPG), designed to generate control policies without a precise numerical model. Instead, it uses an MDP with the action decided via an intra-option policy at each state. DDPG is one of the model-free DRL algorithms with actor-critic architecture that is widely used for continuous action space problems.

Similar to the Bellman equation for $Q(s, a)$, the distributional Bellman equation was proposed to calculate the state-action value distribution $Z(s, a)$. By using a set of quantiles to approximate $Z(s, a)$, the quantile regression-based deep deterministic policy gradient (QR-DDPG) is constructed. The option-value function $Q_\Omega(s_t, \omega_t)$, parameterized by ϕ, is used to describe the utility of an option ω_t at state s_t. For the quantile option architecture (QUOTA), it is assumed that there are N quantile functions for estimating the quantile levels and constructing M options.

For a CMP process, the states of the MDP must contain information related to the control objectives and requirements, such as disturbances and tracking target, in order to properly describe the state of a complex semiconductor manufacturing process. At each run t, the CMP process can choose p_t (downward pressure) and v_t (relative velocity) as action information. The reward is a goal-directed computational approach in which an agent learns to perform a task by interacting with an unknown dynamic environment, but it is still constrained with a limit on how much p_t and v_t can change after any given run.

In the experiments, the total number of training episodes is 50 and each consists of 600 runs. The root-mean-square error (RMSE), variance (σ^2) or outputs, mean absolute deviation (MAD), and mean absolute percentage error (MAPE) are used to measure performance. Comparisons were made with dEWMA, DDPG, and TD3. QUOTA-DDPG performed consistently well, but was outperformed by TD3 in MAD and MAPE under shift disturbance and all others in CPU time.

4.3 Strengths & Weaknesses

The primary strength of this approach is the use of an DDPG with option structure for a model-free DRL algorithm, providing an enhancement to the exploration function in an ϵ-greedy policy. This improved exploration allows the agent to collect more rewards, however its weakness is that it is much slower to converge (in absolute time) due to requiring more CPU time.

4.4 Take-Aways & Limitations

The key take-away is the advantage of the option-value function during exploration and the performance gains that result. Even the limitation of slow convergence can be mitigated through the use of faster machines. However, QUOTA-DDPG will need to perform better under all conditions, and take

into account more than just pressure and velocity, to be a universally accepted solution for CMP process control. This challenge will significantly increase the complexity of the model, which will future exacerbate its weakness of long processing time. Faster compute speeds will help alleviate that to some degree, but a more efficient model will also be required.

5 Fab Scheduling

Title: *Semiconductor Fab Scheduling with Self-Supervised and Reinforcement Learning* [Tas+23]; **Authors:** Pierre Tassel, Benjamin Kovacs, Martin Gebser, Konstantin Schekothin, Patrick Stockermann, & Georg Seidel; **Publisher:** IEEE; **Publication:** Proceedings of the Winter Simulation Conference; **Year:** 2023.

5.1 Problem Addressed

Over the years, Moore's Law has pushed the economic and manufacturing limits of the semiconductor industry by approximately doubling the number of transistors per given area every two years. While originally intended for computers and computer-like machines, semiconductor chips are now found in everything from toasters and refrigerators to cars and airplanes. This increased demand has had an ecological impact as well, with semiconductor manufacturing accounting for 1.3-2.0% of the total U.S. electricity consumption and around 20,000 tons of water per day.

5.2 Solution Summary

This work introduces a method to successfully learn to schedule a semiconductor manufacturing facility more efficiently using deep reinforcement and self-supervised learning by proposing the first adaptive scheduling approach to handle complex, continuous, stochastic, dynamic, modern fab models. Following the MDP formulation in the authors' previous work, the state-transition function of this deep RL agent considers sequences of lots rather than a fixed number of actions assigning lots to machines.

A state in the MDP is given by the set of legal lots at any given time. A lot is "legal" if its next operation can be scheduled to a free resource at that time. Each lot is also associated with a set of features that provide additional information regarding its status and requirements. The action space is all legal lots at a given time, which means that the action space is discrete and the number of actions may vary from time to time. The reward function penalizes tardiness of a lot and long cycle times, while rewarding high throughput, weighting the actual reward based on the lot priority.

At each decision time point, the simulator provides information about the current state of the fab and awaits an ordered list of lots (generally based on priority) to dispatch. In SMT2020 ("Semiconductor Manufacturing Testbed," which serves as test data), lots are assigned to a machine using a strategy that is

greedy, as it merely determines advantageous allocation at that time point, rather than backtracking or looking ahead [Kop+20]. The agent policy architecture is a function that maps the current state of the simulator to the ordered set of legal lots to make dispatchment decisions.

A training horizon of six months was used to provide a good balance between a long training runtime and a diversity of events that will be handled by the agent. The model was compared to a Hierarchical CR and Hierarchical FIFO and improved on almost all metrics while only adding a negligible overhead on computational time with an 1.02ms longer decision time per time point on average.

5.3 Strengths & Weaknesses

The key strength of this method is in its novelty, as it shows the potential of a completely new approach and it is promising to see it on a standard test dataset. However, the results tables indicate that the improvement is well within the margin of error, so the claims of improvement are not statistically significant. Furthermore, even a computational overhead of 1.02ms can add up if done often enough, which will translate to increased cost in a fully-loaded fab.

5.4 Take-Aways & Limitations

The key takeaway is the same as the strength: the potential of a completely new approach. Future work should include a time frame longer than six months to ensure that several cycles of products make it all the way through the process. Removing this limitation may reveal either an improvement with statistical significance, or a lack of any improvement at all.

6 Adaptive Dispatching

Title: *Simulation and Deep Reinforcement Learning for Adaptive Dispatching in Semiconductor Manufacturing Systems* [Sak+23]; **Authors:** Ahmed H. Sakr, Ayman Aboelhassan, Soumaya Yacout, & Samuel Bassetto; **Publisher:** Springer; **Publication:** Journal of Intelligent Manufacturing; **Year:** 2023.

6.1 Problem Addressed

The semiconductor industry has both characteristics of discrete manufacturing and process manufacturing. Process manufacturing is characterized by a reliance on recipes for achieving the desired output from a tool, while discrete manufacturing is characterized by (among other things) complex processing sequences, equipment reconfiguration and recalibration, and customization of individual products. The result of all this complexity is the need for dispatching of lots in a manner that maximizes machine utilization and overall factory yield. Note:

The problem addressed here is similar to the one addressed in **Fab Scheduling** [Tas+23] and **Robust Scheduling** [I B+20] above, but relates to wafer fabrication, not die assembly.

6.2 Solution Summary

The solution presented in this paper is based upon a discrete-event simulation (DES) model for a cast study of a real semiconductor manufacturing system using both data-driven and agent-based approaches. The model simulates the various processing aspects that are normally present in a semiconductor fab and employs Deep-Q-Network (DQN) RL. A DQN uses artificial neural networks as a Q-function approximator, which follows the Bellman optimality equation, to predict the expected reward, Q-value.

The input data for the data-driven model is structured to consider the different operational aspects of the fab that the model needs to simulate, such as processing steps, pre/postprocessing tasks, and all of the tools/machines that exist in a typical fab. The two kinds of agents in the model are dispatching and resource allocation agents, and pre-programmed agents. The former learn to make practical decisions for lot dispatching and equipment allocation. The latter execute sets of logical rules to simulate equipment functionality, including downtime. The system state functions are defined considering the collaborative behavior between the agents, and the reward function represents the feedback of the system for how adequate the agents' actions are.

The DQN dispatching agents were trained over 40 iterations with a batch size of 1000 each, which took 75 minutes. Performance was compared to a heuristic rule-based dispatching and showed a reduction in the total non-value added time percentages, which indicates an improvement in overall throughput.

6.3 Strengths & Weaknesses

The primary strength of this approach is the use of dual agents, which learned to cooperate with each other to improve the overall performance of the dispatching system. This seemed to be related to the complex definition of system state and reward functions leading it to almost behave like an actor-critic model. The primary weakness of this paper is the lack of specificity in the results. There were multiple charts, but very few actual numbers to quantify the performance.

6.4 Take-Aways & Limitations

The main take-away from this paper is the benefit of a combined data-driven and agent-based approach that led to strong cooperation between the agents. However, this appears to be at the expense of flexibility and generalization. It seems unlikely that this solution could be easily applied to a different factory, or even to model changes to the existing one. This limitation makes the solution less desirable since calculating the impact of retooling should be one of the main goals of modeling a fab.

7 Conclusion

This paper summarized current RL-related research as it applies to semiconductor manufacturing through a literature review of papers found through targeted searches in $ar\chi iv$, IEEExplore, and the ACM-DL. The five papers were mostly published in a variety of journals, but primarily focused on efficient scheduling and advanced process control, which are the two main challenges in semiconductor manufacturing that are best suited for RL solutions. Key findings indicate current opportunities to leverage RL models, mostly MDPs, in creative ways. All of the solutions showed promise, but were not without limitations and opportunities. As the models improve and compute power increases, these opportunities will only grow.

Further research is needed to further "close the loop" from analysis and prediction to system and process control. For example, automatically adjusting *all* the parameters of a CMP process (e.g., including slurry rate and temperature, too) to ensure the the best results in the shortest amount of time. The current state-of-the-art does not include such closed-loop control, likely due to the fact that the models are still not as skilled as trained engineers and compute power has historically not been able handle the load. Now that compute power is unlikely to be the bottleneck, more-robust RL agents should be empowered to make real-time changes to input and control parameters, leading to significant improvement in all areas of semiconductor manufacturing.

References

[Cla68] Arthur C. Clarke. *Arthur C. Clarke Quote*. 1968. URL: https://www.azquotes.com/quote/57374. (Accessed: 3.15.2025).

[I B+20] I. B. Park et al. "A Reinforcement Learning Approach to Robust Scheduling of Semiconductor Manufacturing Facilities". In: *IEEE Transactions on Automation Science and Engineering* 17.3 (July 2020), pp. 1420–1431. ISSN: 1558-3783. DOI: 10.1109/TASE.2019.2956762.

[Kop+20] Denny Kopp et al. "SMT2020—A Semiconductor Manufacturing Testbed". In: *IEEE Transactions on Semiconductor Manufacturing* 33.4 (2020), pp. 522–531. DOI: 10.1109/TSM.2020.3001933.

[Cho+21] Hyung-Min Cho et al. "A Chemical Monitoring and Prediction System in Semiconductor Manufacturing Process Using Bigdata and AI Techniques". In: *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. Apr. 2021, pp. 488–491. DOI: 10.1109/ICAIIC51459.2021.9415241.

[LDJ21] Yanrong Li, Juan Du, and Wei Jiang. "Reinforcement Learning for Process Control with Application in Semiconductor Manufacturing". In: (2021). arXiv: 2110.11572 [eess.SY]. URL: https://arxiv.org/abs/2110.11572.

[Kov+23] Benjamin Kovács et al. "A Customizable Reinforcement Learning Environment for Semiconductor Fab Simulation". In: *Proceedings of the Winter Simulation Conference*. WSC '22. Place: Singapore, Singapore. IEEE Press, 2023, pp. 2663–2674.

[MP23] Zhu Ma and Tianhong Pan. "Distributional reinforcement learning for run-to-run control in semiconductor manufacturing processes". In: *Neural Computing and Applications* 35.26 (Sept. 2023), pp. 19337–19350. ISSN: 1433-3058. DOI: 10.1007/s00521-023-08760-1. URL: https://doi.org/10.1007/s00521-023-08760-1.

[Sak+23] Ahmed H. Sakr et al. "Simulation and deep reinforcement learning for adaptive dispatching in semiconductor manufacturing systems". In: *Journal of Intelligent Manufacturing* 34.3 (Mar. 2023), pp. 1311–1324. ISSN: 1572-8145. DOI: 10.1007/s10845-021-01851-7. URL: https://doi.org/10.1007/s10845-021-01851-7.

[Tas+23] Pierre Tassel et al. "Semiconductor Fab Scheduling with Self-Supervised and Reinforcement Learning". In: (2023). arXiv: 2302.07162 [cs.AI]. URL: https://arxiv.org/abs/2302.07162.