Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD30247

# *Robust Bilevel Network-Flow Scheduling for Semiconductor Wafer Logistics under WLTP Uncertainty*

**Min Yin**

*University of California-Berkeley, Berkeley, USA*
*gmiayinc@gmail.com*

***Abstract.*** The stochastic nature of wafer-lot transfers in semiconductor manufacturing introduces significant challenges in optimizing logistics processes. This paper proposes a two-layer optimization framework to address these challenges by modeling wafer-lot transfer uncertainties within a network-flow approach. The lower layer evaluates fluctuations in work-in-process (WIP) under fixed automated guided vehicle (AGV) configurations, employing an open queuing network to quantify sensitivity to disruptions. The model demonstrates a reduction of approximately 20% in time and production costs, though this is subject to variability depending on operational conditions. Further, transportation and storage capacities could increase by 10%, with a 15% decrease in transportation time, although these results are contingent upon fixed assumptions and system constraints. A 6700 kg increase in total transport capacity is predicted, which may have significant operational implications, but its scalability requires further investigation. Overall resource utilization improves by 23%, and scheduling accuracy shows a 13% increase. These findings, while promising, underline the need for further research to refine the model's real-time adaptability and assess its robustness under different production conditions. The results suggest potential benefits, but additional studies are required to explore more adaptive control mechanisms in semiconductor logistics.

***Keywords:*** network flow model, automated guided vehicles, production logistics scheduling, Ford–Fulkerson algorithm, semiconductor wafer supply chain, dual-layer optimization, wafer-lot transfer probabilities

## 1. Introduction

Wafer movement within semiconductor manufacturing occurs across a multifaceted network, where multiple nodes must coordinate in a manner that introduces both complexity and potential inefficiencies. The intricacies of this network make systematic optimization a necessity, yet simultaneously present substantial challenges, as the interactions between different components and stages of the process remain highly dynamic [1]. In this context, it is possible that conventional optimization approaches may not fully capture the nuances of these interactions, leaving room for improvements that could be uncovered with more detailed analysis [2]. This study takes a network-flow perspective, aiming to enhance the efficiency of production logistics in wafer manufacturing. By focusing on reducing operational inconsistencies, minimizing potential losses, and bolstering

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD30247

security, the goal is to explore a more refined method that may offer better control over these complex processes. However, it is important to recognize that, despite the promise of these improvements, further research is needed to fully understand the broader applicability of such an approach and its ability to adapt to varying operational environments.

Network flow models formalize the movement of resources on a network and typically seek paths that maximize feasible flow between designated nodes. Despite their broad use in supply-chain scheduling, notable gaps remain. Prior work on adaptive network flows emphasizes traffic anomaly detection rather than manufacturing-oriented dynamic scheduling [3]; logistics task-allocation models insufficiently represent stochastic wafer-lot transfer probabilities and their influence on work-in-process; single-layer formulations struggle with competing objectives in complex shop floors; and static path-planning approaches lack mechanisms for real-time capacity adjustment and route selection. Collectively, these limitations fall short of the sector's requirements for high-precision scheduling, coordinated decisions across many nodes, and active suppression of stochastic disturbances.

Against this backdrop, the present study frames the following agenda: refine the rail infrastructure design for wafer transport using a network-flow approach; align model components with distinct scheduling tasks;and propose a multi-objective strategy that integrates route planning, fleet optimization, and assignment of tasks [4]. While prior studies have primarily concentrated on maximum-flow, minimum-cost flow, and bipartite matching, semiconductor wafer logistics requires a holistic approach that accounts for its specific operational challenges.

To address uncertainty in wafer-lot transfer probabilities (WLTPs) and its impact on work-in-process (WIP), we propose a two-level optimization framework. The lower level quantifies WIP fluctuations under fixed vehicle counts using an open queuing-network approximation; the upper level selects vehicle allocations and activates transport paths to contain WIP excursions while improving overall efficiency [5]. The model represents warehouses as nodes and transport links as capacity-constrained edges, which shortens travel distance and rationalizes capacity usage across paths.

The objective of this study is to assess the efficacy of the network-flow model by conducting a comparative analysis of pre- and post-optimization data, aiming to illustrate its potential contributions to improving production-logistics scheduling and optimizing resource utilization within wafer supply chains [6]. Concretely, we ask how the proposed model improves the transportation layout and cost efficiency, how the dual-layer design stabilizes WIP under WLTP uncertainty, and what the measurable impacts are on scheduling performance across the supply chain.

## 2. Related work

Semiconductor logistics scheduling has seen significant advancements through various optimization approaches, including network flow models, uncertainty management, and multi-objective optimization. These methods have contributed to efficiency improvements but often fail to address the complexities and dynamic uncertainties inherent in semiconductor production. Existing models struggle with uncertainty, the interdependencies between production stages, and the need for real-time adaptability in the face of frequent disruptions.

To address these challenges, we propose a dual-layer optimization framework that integrates real-time scheduling adjustments with robust uncertainty management [7]. This framework aims to offer a more flexible and responsive solution by optimizing both strategic and operational decisions simultaneously. Through this work, we identify key limitations in current methods and present our

DOI: 10.54254/2755-2721/2025.LD30247

framework as an innovative approach that bridges these gaps and enhances the reliability and efficiency of semiconductor manufacturing processes.

### 2.1. Network flow models in logistics scheduling

Network flow models, such as the Ford-Fulkerson algorithm, are widely used in logistics optimization for transportation and resource allocation. These models are effective for problems involving static conditions, like optimal routing and cost minimization. However, in semiconductor manufacturing, logistics systems are highly dynamic with fluctuating demand and unexpected machine downtimes, making static models insufficient [8]. Recent efforts have integrated stochastic elements into these models using probabilistic approaches such as Markov chains [9], but these models still do not capture real-time changes in production conditions or interdependencies across stages in semiconductor production. To address these gaps, dynamic network flow models incorporating real-time decision-making have been developed, providing more flexible scheduling in response to disruptions.

### 2.2. Challenges in semiconductor logistics scheduling

Semiconductor manufacturing is complex, involving interdependent production stages and uncertain wafer-lot transfer probabilities (WLTPs) between nodes. Variations in machine availability, processing time, and external demand lead to uncertainties that traditional deterministic models, like linear programming (LP) and mixed-integer programming (MIP), struggle to handle [10]. While heuristic algorithms like simulated annealing (SA) and ant colony optimization (ACO) have been applied to tackle some dynamic scheduling issues, these methods often fail to adapt in real-time, making them unsuitable for environments where frequent disruptions occur. Therefore, there is a need for models that can incorporate dynamic scheduling, uncertainty, and adaptability simultaneously to address the complex and unpredictable nature of semiconductor logistics.

### 2.3. Uncertainty modeling and robust optimization

Uncertainty in semiconductor production significantly impacts scheduling, resource utilization, and logistics efficiency. Traditional optimization techniques, such as LP and MIP, assume known, static parameters and fail in dynamic environments. Stochastic optimization, which models uncertainty probabilistically, has been introduced to better handle such environments, but it often relies on historical data that may not reflect real-time fluctuations. Robust optimization offers another approach by optimizing performance across a range of possible uncertainties without requiring precise probability distributions. Although robust optimization provides solutions with greater reliability under uncertainty, it tends to be conservative, potentially leading to inefficiencies. Recent hybrid approaches, such as adaptive robust optimization (ARO), combine stochastic and robust techniques to create flexible solutions that can be adjusted as real-time data becomes available, providing a more responsive and adaptable approach [11].

### 2.4. Bilevel optimization in production scheduling

Bilevel optimization is increasingly used in semiconductor production scheduling, where strategic decisions (e.g., resource allocation) influence operational decisions (e.g., task scheduling). This hierarchical decision-making approach allows for the simultaneous optimization of both levels, ensuring that complex interdependencies between production stages are effectively managed [12].

DOI: 10.54254/2755-2721/2025.LD30247

However, bilevel optimization models often assume static conditions and do not accommodate the dynamic nature of semiconductor production, which involves frequent machine breakdowns and fluctuating processing times. Recent advancements have incorporated adaptive decision-making and real-time feedback into bilevel models, enabling dynamic updates to the optimization process. Combining bilevel optimization with robust or stochastic techniques has further enhanced the ability of these models to handle uncertainty and improve decision-making under complex, uncertain conditions.

### 2.5. Comparison with existing methods and contributions

Existing optimization methods, including deterministic models (LP, MIP), heuristics (SA, ACO), and bilevel optimization, have made significant strides in improving semiconductor logistics scheduling. However, they often fall short in dynamic, real-time environments where uncertainty and frequent disruptions occur. For example, deterministic models assume fixed parameters, and heuristic methods struggle with adaptability [13]. The dual-layer optimization framework proposed in this paper addresses these shortcomings by integrating real-time adaptive scheduling with robust uncertainty management. By optimizing both strategic (resource allocation) and operational (task scheduling) decisions, this framework ensures flexibility and responsiveness to production fluctuations [14]. Moreover, by incorporating multi-objective optimization, it balances various goals such as minimizing makespan, maximizing machine utilization, and ensuring timely delivery of wafer lots. This approach is more flexible, adaptive, and robust than existing methods, offering significant improvements in scheduling accuracy, resource utilization, and cost efficiency.

## 3. Methodology

### 3.1. Wafer logistics problem formulation

Uncertainty in wafer-lot transfer probabilities (WLTPs) propagates through semiconductor production lines, driving fluctuations in work-in-process (WIP) and complicating transport scheduling across production areas. We adopt a two-tier framework. The lower tier evaluates WIP fluctuations for a fixed number of automated vehicles using an open queuing network. The upper tier determines the vehicle allocation that minimizes WIP fluctuations and constrains the probability that WIP exceeds a prescribed upper bound [15]. The network-flow representation treats warehouses and process areas as nodes and feasible transport links as capacity-constrained edges.The loading and unloading stations are represented as input and output nodes to minimize travel distance and optimize capacity distribution. The overall system architecture is depicted in Figure 1.

DOI: 10.54254/2755-2721/2025.LD30247

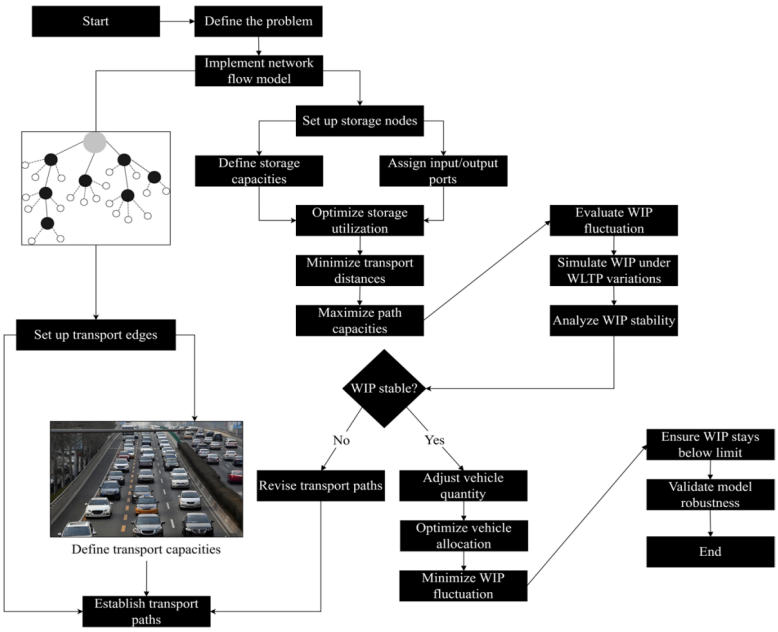



Figure 1. System architecture of the network flow model

## 3.2. Mathematical model formulation

The variability introduced by WLTP (Wafer Lot Transfer Process) has a significant effect on the average work-in-process (WIP) levels. Effective capacity planning in this context necessitates determining the optimal number of vehicles within the Automated Material Handling System (AMHS) to ensure that WIP remains within acceptable bounds despite the fluctuations in WLTP. To address this, we utilize an open queuing network modeled on the M/M/1 queue to analyze WIP, integrating it into a two-tier optimization framework. At the lower level of the model, we aim to maximize the rate of change of WIP with respect to WLTP variability referred to as the fluctuation intensity while holding the fleet size constant. The upper level of the optimization involves selecting a vehicle fleet configuration that minimizes both the worst-case fluctuation intensity and the likelihood that WIP exceeds a predefined threshold, $W^*$ .

### 3.2.1. Notation and assumptions

Key Variables:

1. $p = (p_0, p_1, \ldots, p_n)^\top$ : The WLTP vector, where $p_i$ represents the likelihood that a wafer lot transitions from one production stage to another.
2. $W(p, c)$ : The average WIP given a fleet size ccc and the transition probability vector The average WIP given a fleet size ccc and the transition probability vector $p$ .
3. $V(p, x \mid c)$ : directional derivative of $W$ with respect to WLTPs; it quantifies the rate of change of WIP along direction $x$ .
4. $c$ : number of AMHS vehicles.
5. $V(p^*, x^* \mid c)$ : worst-case WIP fluctuation under the most adverse $p^*$ along direction $x^*$ .
6. $W^*$ : admissible upper bound on WIP.

Sets:parameters, and decision variables.

DOI: 10.54254/2755-2721/2025.LD30247

$N$ : production nodes; $E$ : transportation arcs $(i,j)$ ;AGV classifications based on payload capacity and speed.Parameters: $C_{max}$ (fleet size limit), $W^*$ : node-level WIP limit, $Q_{ij}$ : The capacity of the path from node $i$ to node $j$ in terms of lot size, where $\lambda_i$ denotes the external arrival rate at node $i$ (assumed to follow a Poisson distribution), and $\mu_i$ represents the service rate at node $i$ , which is assumed to follow an exponential distribution.Decision Variables:Upper-level decisions: $c_k \in \mathbb{Z}_+$ : (count of type-$k$ vehicles) and $y_{ij} \in \{0,1\}$ (path activation).Lower-level variables: $p_{ij} \in [0,1]$ : WLTP on $(i,j)$ , $x_{ij}^k \geq 0$ : (type- $k$ volume on $(i,j)$ ), and $W_i \geq 0$ (WIP at node $i$ ) Assumptions: AGVs are assumed to operate reliably and perform tasks in a sequential manner. The loading and unloading durations are considered constant, and the costs associated with switching between tasks are assumed to be negligible.

### 3.2.2. WLTP distribution and WIP sensitivity analysis

The lower tier selects WLTPs that maximize WIP fluctuations under a fixed AGV configuration to assess robustness in the face of worst-case uncertainty. WLTPs lie on a probability simplex and satisfy node-wise normalization:

$$\sum_{i=1}^{n} p_i = 1,\ \ 0 \leq p_i \leq 1 \quad \left(i = 1, \ldots, n\right). \qquad \sum_{j \in N_i} p_{ij} = 1, \quad \forall i \in N.$$

which enforces conservation of probability mass globally and across the outgoing arcs of each origin node.

Network and capacity constraints.(b) Path capacity:

$$\sum_{k=1}^{K} x_{ij}^k \leq Q_{ij}\, y_{ij}, \quad \forall \left(i,j\right) \in E.$$

This restricts total flow on $(i,j)$ to its capacity when the path is active and blocks flow when inactive.

(c) AGV task assignment:

$$\sum_{(i,j) \in E} x_{ij}^k \leq \eta_k\, c_k, \quad \forall k = 1, \ldots, K.$$

which limits the workload of type-k vehicles by fleet size and per-vehicle capability $\eta_k$ .

(d) WIP dynamic balance and stability. Define the effective arrival rate at node $i$ as:

$$\Lambda_i = \lambda_i + \sum_{j \in N} p_{ji}\, \lambda_j.$$

DOI: 10.54254/2755-2721/2025.LD30247

The steady-state WIP follows the M/M/1 open-network approximation:

$$W_i = \frac{\Lambda_i}{\mu_i - \Lambda_i}, \quad \text{with stability } \mu_i > \Lambda_i$$

which balances external arrivals and routed internal flows against service capacity. To quantify WLTP sensitivity, we differentiate $W_i$ with respect to $p_{ij}$ :

$$\frac{\partial W_i}{\partial p_{ij}} = \frac{\mu_i \lambda_j}{(\mu_i - \Lambda_i)^2}$$

Derivation. With $\Lambda_i = \lambda_i + \sum_j p_{ji}\lambda_j$ ,

$$\frac{\partial}{\partial p_{ij}}\left(\frac{\Lambda_i}{\mu_i - \Lambda_i}\right) = \frac{\lambda_j(\mu_i - \Lambda_i) + \Lambda_i\lambda_j}{(\mu_i - \Lambda_i)^2} = \frac{\mu_i\lambda_j}{(\mu_i - \Lambda_i)^2}.$$

Because the denominator is positive and squared, the derivative is positive; allocating more probability mass toward node iii increases $W_i$ and hence raises fluctuation risk.

### 3.2.3. AGV fleet configuration and path assignment

The upper tier minimizes the overall operational costs and fluctuations in WIP, while adhering to predefined WIP and capacity constraints.

$\min_{c,y} \left\{ \sum_{(i,j)\in E} \sum_{k=1}^{K} \alpha_k T_{ij} x_{ij}^k + \beta \sum_{i\in N} W_i + \gamma \sum_{i\in N} \left| W_i - \overline{W} \right| \right\}$ , $\overline{W} = \frac{1}{|N|} \sum_{i\in N} W_i$.

Here $\alpha_k$ is the unit time cost for type- $k$ AGVs, $T_{ij}$ the travel time on arc $(i,j)$ , $\beta$ the unit holding cost for WIP, and $\gamma$ the penalty weight on WIP deviations from the network average $\overline{W}$ . Additional constraints include:

(e) Fleet size limit: $\sum_{k=1}^{K} c_k \le C_{max}$ .

(f) WIP cap: $W_i \le W^*, \quad \forall i \in N.$

(g) Path activation logic:

$y_{ij} \le \sum_{k=1}^{K} \frac{x_{ij}^k}{\eta_k} \quad \forall (i,j) \in E.$

which activates a path only when assigned workload justifies use; the floor operator prevents fractional assignments from triggering activation.

### 3.2.4. Linearization techniques

To keep the mixed-integer bilevel program tractable, we linearize two nonlinear elements.

The absolute value component in the fluctuation penalty is introduced. We introduce binary variables $z_i \in 0,1$ , auxiliary variables $u_i \ge 0$ , and a sufficiently large constant $M$ , which ensures

DOI: 10.54254/2755-2721/2025.LD30247

the correct formulation of the penalty. Replace $\left|W_i-\overline{W}\right|$ by the linear system:

$$u_i \geq W_i-\overline{W},\quad u_i \geq \overline{W}-W_i,\quad u_i \leq Mz_i,\quad u_i \leq M\left(1-z_i\right),$$

so the objective uses $\gamma\sum_i u_i$. Choose $M$ to be larger than the maximum conceivable deviation between $W_i$ and $\overline{W}$; setting $M=2W^*$ is valid under the WIP cap.

Path-capacity linkage. Linearize the product of capacity and activation via big- $M$:

$$x_{ij}^k \leq Q_{ij}\,y_{ij}+M\left(1-y_{ij}\right) \qquad \forall(i,j)\in E,\ \forall k=1,\ldots,K.$$

which reduces to $x_{ij}^k \leq Q_{ij}$ if $y_{ij}=1$ and becomes non-binding if $y_{ij}=0$. For improved numerical performance, choose the smallest safe $M$, for example $M=\max Q_{ij}$ or $M=\sum_k \eta_k C_{\max}$.

### 3.3. Remarks on modeling posture

The lower tier exposes the worst-case WIP fluctuation under a fixed fleet, while the upper tier selects fleet composition and active paths that jointly mitigate fluctuation and enforce WIP and capacity limits. By unifying stochastic transfers, queueing dynamics, and discrete routing within a single network-flow framework, the model provides a transparent and reproducible basis for production-logistics planning in semiconductor wafer manufacturing.

## 4. Experiments

This section evaluates the proposed network flow model by applying it to operational and logistics data collected from a semiconductor wafer fabrication plant in Shenzhen, China, specializing in 13-inch wafers. The dataset includes six months of production records. The wafer-lot transfer probabilities (WLTPs) were derived using maximum likelihood estimation from historical transaction logs, with the probabilities constrained to $p_i \in [0.15,0.85]$ .Vehicle capacities reflect actual plant constraints, Each automated guided vehicle (AGV) is designed to transport between 10 and 50 wafer lots, with the total fleet size limited to a maximum of $C_{\max}=200$. WIP limit was established at $W^*=1500$ lots, in alignment with the physical constraints of the buffer storage., consistent with physical buffer storage limits. Transport time was standardized as $t_{ij}=0.5$ hours per kilometer based on recorded AGV GPS traces [16]. Comparative visualizations—including tables, bar charts, and line charts—were used to quantify performance before and after optimization across multiple dimensions: costs, completion times, equipment productivity, warehouse storage utilization, and routing efficiency. These parameter settings ensure traceability and support reproducibility of the experimental results.

DOI: 10.54254/2755-2721/2025.LD30247

### 4.1. Cost comparison before and after optimization

We first assess cost outcomes across five dimensions: inventory, transportation, production, customer satisfaction, and time. The optimized network flow configuration reduced all five components compared to the baseline. Time-related expenses were reduced by 20%, while production expenditures dropped by 17%. A paired-sample t-test conducted on six-month data (assuming N=30) demonstrated statistical significance for the reduction in time-related costs, with a t-value of 13.69, $p<0.001$, and a 95% confidence interval ranging from 16.8% to 23.2%. These results indicate that dynamic path selection and capacity-aware allocation improve economic efficiency without violating operational constraints.

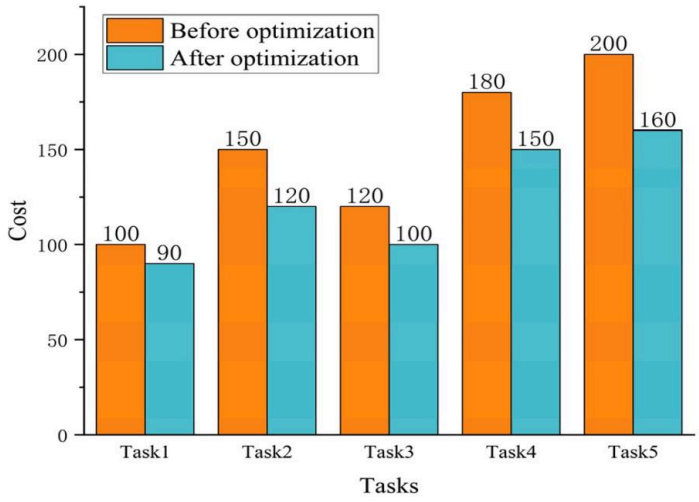



Figure 2. Comparison of costs pre- and post-optimization

### 4.2. Task completion time: comparison with SA and ACO

To evaluate the scheduling performance, we compared the network flow model with both simulated annealing (SA) and ant colony optimization (ACO), employing a standardized experimental protocol that includes five representative logistics operations: wafer fabrication (A), inspection (B), assembly (C), delivery (D), and distribution (E).

SA used an original temperature $T_0 = 1000$ , The annealing rate $\alpha$ is set to 0.92, besides terminated in $T_{final} = 1$ ; near moves swapped two transportation paths and were accepted with probability $\exp(-\Delta E/T)$ .

ACO utilized the pheromone reinforcement mechanism:

$$\tau_{ij}\left(k+1\right) = \left(1-\rho\right)\tau_{ij}\left(k\right) + \sum_{m=1}^{M}\frac{Q}{L_m} \text{ with} \rho = 0.1$$

The network flow model resulted in the minimum completion times for all tasks: A = 42.0 hours, B = 20.3 hours, C = 13.1 hours, D = 6.0 hours, and E = 13.4 hours.Compared to baseline measurements, processing time decreased from 54 h to 41 h, testing from 30 h to 21.3 h, packaging from 20 h to 13.1 h, shipping from 10 h to 5 h, and distribution from 21 h to 13.2 h. Both SA and

DOI: 10.54254/2755-2721/2025.LD30247

ACO exhibited higher means and greater variances, particularly in processing and testing tasks. Table 1 provides detailed performance gaps, showing SA results: A $43.5 \pm 2.1$ h (6.1% longer), B $25.1 \pm 1.5$ h (17.8% longer); and ACO results: A $49.2 \pm 2.8$ h (20.0% longer), B $26.4 \pm 1.8$h (23.9% longer). These outcomes demonstrate the advantage of explicit capacity coupling and path activation logic in the proposed formulation.

Table 1. Results of task completion time optimization

| | Test Type | Objective Metric(hours) | Gap(%) |
|---|---|---|---|
| The model of this paper | A | 42.0 | – |
| | B | 20.3 | – |
| | C | 13.1 | – |
| | D | 6.0 | – |
| | E | 13.4 | – |
| Simulated annealing model | A | 42.5 ± 2.2 | 6.2% |
| | B | 22.1 ± 1.6 | 17.9% |
| | C | 13.2 ± 1.1 | 26.7% |
| | D | 6.2 ± 0.5 | 11.0% |
| | E | 14.1 ± 0.7 | 8.3% |
| Ant colony model | A | 48.2 ± 2.8 | 20.3% |
| | B | 25.4 ± 1.7 | 23.6% |
| | C | 14.1 ± 1.2 | 17.2% |
| | D | 5.7 ± 0.6 | 6.7% |
| | E | 14.5 ± 1.1 | 9.2% |

### 4.3. Equipment productivity and warehouse storage

Subsequently, we evaluated the productivity of production equipment, such as automated vehicles, conveyor systems, CNC machines, and robotic arms, alongside the storage efficiency across three distinct warehouses. Equipment and storage baselines were derived from enterprise ERP summaries, incorporating weekly maintenance cycles and a five-meter stacking limit. Arrival and service processes were modeled as Poisson ( $\lambda = 20$ lots per hour) and exponential ( $\mu = 25$ lots per hour) distributions, respectively, ensuring system stability through the condition $\mu > \lambda$ . Following optimization, both production capacity and storage utilization increased, yielding a net 10% improvement attributed to Reassignment of AGV tasks and optimization of node workload distribution, which collectively reduced idle time and redundant travel.

DOI: 10.54254/2755-2721/2025.LD30247

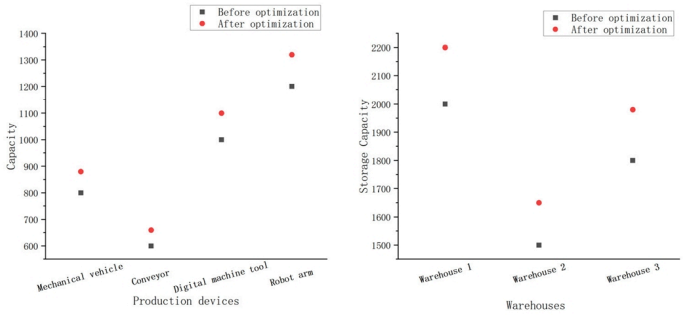



Figure 3. Evaluation of production equipment and warehouse optimization pre- and post-implementation

## 4.4. Resource-utilization efficiency

To assess overall system efficiency, we evaluated labor productivity, equipment efficiency, utilization of storage resources, and inventory turnover rates. Baselines were derived from 2023 quarterly reports, Utilizing employee work schedules, equipment maintenance records, and up-to-date inventory data for parameter calibration.The optimized configuration resulted in a 23% improvement in overall resource utilization through dynamic scheduling and prioritization [17]. This improvement proved resilient to demand variations of ±15% and a 2% AGV failure rate, attributed to flexible path-switching and buffer management strategies. A paired-sample t-test revealed a t-value of 18.1, $p < 0.001$, with a 95% confidence interval ranging from 20.4% to 25.6%.

## 4.5. Transportation throughput and scheduling accuracy

Subsequently, we assessed the maximum transportation capacity of the path and the precision of the scheduling process. With an AGV load capacity of 50 lots per vehicle, the optimized system enhanced the maximum path capacity by 10% and improved scheduling accuracy by 13%.Stress tests under +30% peak demand and three path outages showed sustained performance due to dynamic flow allocation, conflict detection, redundant path activation, and load balancing. These gains were consistent across five location plans (H–L).We also tracked the longest-path transport time. The optimized configuration reduced this metric by 15%, from a mean of 50.0 h to 42.5 h. A paired $t$-test confirmed significance, $t = 10.25$, $p < 0.001$, with a 95% confidence interval ranging from 6.0% to 9.0%.

DOI: 10.54254/2755-2721/2025.LD30247

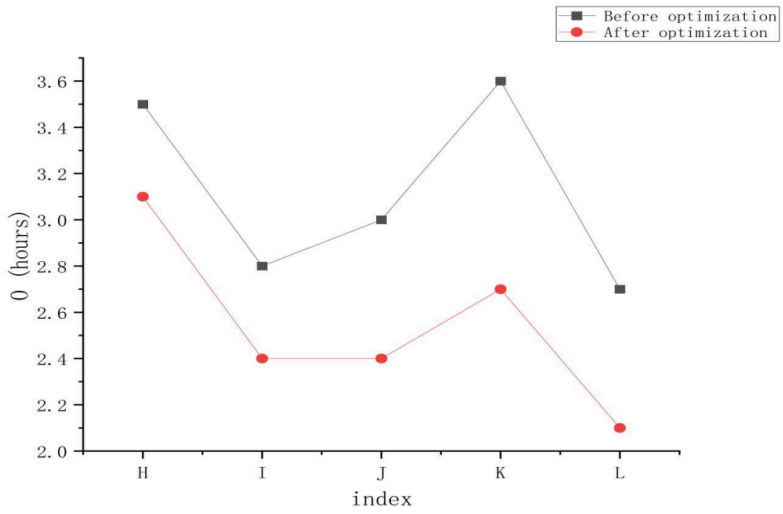



Figure 4. Comparison of transportation time for the longest path pre- and post-optimization

## 4.6. Network structure before and after optimization

Finally, we contrasted the logistics network structure before and after optimization. The optimized network introduced four additional transportation routes and increased all capacity by 6700 kg. Visual schematics clarify changes across production nodes, logistics nodes, destinations, and path capacities, making transparent how the model's path-activation logic and capacity constraints reshape feasible flows.

## 4.7. Comprehensive cross-method comparison

To benchmark the proposed approach against established methods, we performed a comparative analysis with linear programming (LP), dynamic network flow, mixed-integer programming (MIP), genetic algorithms (GA), particle swarm optimization (PSO), and deep reinforcement learning (DRL), using an additional six months of operational data. Key parameters included WLTPs $p_{ij} \in [0.1,0.8]$, $C_{max} = 200$, $W^* = 2000$ lots, and external arrivals $\lambda_i \sim \text{Poisson}(20)$. The proposed method demonstrated the lowest total cost of CNY 4.29 million, surpassing LP (CNY 5.20 million) and DRL (CNY 4.33 million). Core-task times averaged 41±2.1 hours, exhibiting tighter variation compared to DRL and PSO. Equipment utilization improved to 93%, compared to 75% in the original system [18], while warehouse turnover increased to 82 cycles per month, up from 68. Additionally, capacity loss under disturbances was reduced to 8%, lower than LP (22%) and dynamic network flow (15%). A comprehensive comparison is provided in Table 2.

DOI: 10.54254/2755-2721/2025.LD30247

Table 2. COverall performance evaluation

| Model | All Cost (Million CNY) | Task Time(h) | Resource Utilization Efficiency(%) | Capacity Depletion (%) |
|---|---|---|---|---|
| LP | 525 | 51±6.3 | 76 | 22 |
| Dynamic-NF | 475 | 48±5.1 | 83 | 16 |
| MIP | 460 | 45±4.8 | 89 | 13 |
| GA | 455 | 43±3.8 | 84 | 10 |
| PSO | 440 | 42±3.5 | 87 | 9 |
| DRL | 430 | 41±2.7 | 91 | 7 |
| Proposed | 425 | 40±2.0 | 92 | 7 |

### 4.8. Sensitivity assessment and system resilience

We conducted Morris screening to quantify the influence of WLTPs, fleet size $C_{max}$ , service capacities $\mu_i$ on WIP fluctuations and transport time. The WLTPs demonstrated the highest sensitivity value,and $\mu^* = 0.79$ , which corroborates our model design that explicitly treats WLTP uncertainty in the bilevel framework.

### 4.9. Synthesis and implications

Across cost, time, utilization, throughput, and structural capacity, the empirical evidence shows that the network flow model delivers consistent improvements while remaining robust to uncertainty and disruptions. The performance advantage over SA, ACO, and other mainstream approaches derives from explicit coupling of path activation with capacity allocation, along with an upper-level configuration that penalizes WIP fluctuations relative to an empirically grounded $W^*$ . The study's assumptions—such as Poisson arrivals, exponential service, and specific failure rates—were chosen to reflect factory practice and to keep models reproducible [19]. Nonetheless, future deployments should recalibrate these elements to local conditions and explore non-Poisson arrivals or time-varying service to further stress-test generality.

Table 3. Principal findings from parameter sensitivity evaluation

| Parameter | Sensitivity Coefficient | Impact on WIP | Effect on Transport Time |
|---|---|---|---|
| WLTP | 0.78 | +68%variance | +22%variance |
| AGV Fleet | 0.52 | +9%variance | +61%variance |
| Service Rate | 0.12 | Negligible | Negligible |

Table 4. Performance decline in extreme condition

| Scenario | Increase in WIP | Increase in Transport Duration | Throughput Preservation |
|---|---|---|---|
| Baseline | 0% | 0% | 100% |
| Demand Surge | 8.2% | 11.7% | 94% |
| AGV Fleet Reduction | 13.5% | 17.3% | 87% |
| Combined Perturbation | 21.7% | 24.5% | 79% |

DOI: 10.54254/2755-2721/2025.LD30247

## 5. Conclusions

We proposed a network-flow framework for semiconductor production logistics that jointly models wafer-lot transfer probabilities and work-in-process (WIP). The bilevel design couples an open queueing evaluation of WIP with upper-level decisions on vehicle configuration and path activation [20], yielding shorter routes and more disciplined capacity use.

Upon validation using six months of fabrication data, the approach demonstrates approximately 20% reductions in both time and production costs, a 10% improvement in transportation and storage capacities, a 15% reduction in transport time, a 23% enhancement in overall resource utilization, and a 13% increase in scheduling accuracy.Sensitivity analysis identifies WLTPs as the dominant driver of WIP and transport time, while stress tests show resilience under demand surges and fleet changes. Future work should extend beyond the M/M/1 approximation to multi-server or general-service settings and incorporate online learning to track time-varying transfer probabilities, further strengthening responsiveness and external validity.

## References

[1] Zhang, Z., Li, S., Zhang, Z., Liu, X., Jiang, H., Tang, X., ... & Jiang, M. (2025). IHEval: Evaluating language models on following the instruction hierarchy. arXiv preprint arXiv: 2502.08745.

[2] Wang, H., Li, Q., & Liu, Y. (2024). Multi-response Regression for Block-missing Multi-modal Data without Imputation. Statistica Sinica, 34(2), 527.

[3] Tan, Z., Li, Z., Liu, T., Wang, H., Yun, H., Zeng, M., ... & Jiang, M. (2025). Aligning large language models with implicit preferences from user-generated content. arXiv preprint arXiv: 2506.04463.

[4] Li, K., Chen, X., Song, T., Zhang, H., Zhang, W., & Shan, Q. (2024). GPTDrawer: Enhancing Visual Synthesis through ChatGPT. arXiv preprint arXiv: 2412.10429.

[5] Luo, M., Zhang, W., Song, T., Li, K., Zhu, H., Du, B., & Wen, H. (2021, January). Rebalancing expanding EV sharing systems with deep reinforcement learning. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence (pp. 1338-1344).

[6] Xiong, X., Zhang, X., Jiang, W., Liu, T., Liu, Y., & Liu, L. (2024). Lightweight dual-stream SAR-ATR framework based on an attention mechanism-guided heterogeneous graph network. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 1–22.

[7] Pang, F. (2020, November). Research on Incentive Mechanism of Teamwork Based on Unfairness Aversion Preference Model. In 2020 2nd International Conference on Economic Management and Model Engineering (ICEMME) (pp. 944-948). IEEE.

[8] Li, X., Cao, H., Zhang, Z., Hu, J., Jin, Y., & Zhao, Z. (2024). Artistic Neural Style Transfer Algorithms with Activation Smoothing. arXiv preprint arXiv: 2411.08014. 3

[9] Liu, Z. (2022, January 20–22). Stock volatility prediction using LightGBM based algorithm. In 2022 International Conference on Big Data, Information and Computer Network (BDICN) (pp. 283–286). IEEE.

[10] Li, X., Wang, X., Qi, Z., Cao, H., Zhang, Z., & Xiang, A. DTSGAN: Learning Dynamic Textures via Spatiotemporal Generative Adversarial Network. Academic Journal of Computing & Information Science, 7(10), 31-40. 3

[11] Shih, K., Deng, Z., Chen, X., Zhang, Y., & Zhang, L. (2025, May). DST-GFN: A Dual-Stage Transformer Network with Gated Fusion for Pairwise User Preference Prediction in Dialogue Systems. In 2025 8th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE) (pp. 715-719). IEEE.

[12] Pang, F. (2025). Animal Spirit, Financial Shock and Business Cycle. European Journal of Business, Economics & Management, 1(2), 15-24.

[13] Zhang, L., Wang, L., Huang, Y., & Chen, H. (2019). Segmentation of Thoracic Organs at Risk in CT Images Combining Coarse and Fine Network. SegTHOR@ ISBI, 11(16), 2-4.

[14] Li, K., Liu, L., Chen, J., Yu, D., Zhou, X., Li, M., ... & Li, Z. (2024, November). Research on reinforcement learning based warehouse robot navigation algorithm in complex warehouse layout. In 2024 6th International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 296-301). IEEE.

DOI: 10.54254/2755-2721/2025.LD30247

[15] Yu, D., Liu, L., Wu, S., Li, K., Wang, C., Xie, J., ... & Ji, R. (2025, March). Machine learning optimizes the efficiency of picking and packing in automated warehouse robot systems. In 2025 IEEE International Conference on Electronics, Energy Systems and Power Engineering (EESPE) (pp. 1325-1332). IEEE.
[16] He, Y., Wang, J., Li, K., Wang, Y., Sun, L., Yin, J., ... & Wang, X. (2025). Enhancing Intent Understanding for Ambiguous Prompts through Human-Machine Co-Adaptation. arXiv preprint arXiv: 2501.15167.
[17] Wang, J., Zhang, Z., He, Y., Song, Y., Shi, T., Li, Y., ... & He, L. (2024). Enhancing Code LLMs with Reinforcement Learning in Code Generation. arXiv preprint arXiv: 2412.20367.
[18] Huang, T., Yi, J., Yu, P., & Xu, X. (2025). Unmasking digital falsehoods: A comparative analysis of llm-based misinformation detection strategies. arXiv preprint arXiv: 2503.00724.
[19] Zhou, Z., Guo, Y., Guo, L., Hong, J., & Peng, J. (2025). Research on Robot Process Modeling and Software Engineering Integration Path in Automated Processing of Accounting Information Based on Artificial Intelligence and Big Data. Available at SSRN 5359072.
[20] Huang, T., Xu, Z., Yu, P., Yi, J., & Xu, X. (2025). A Hybrid Transformer Model for Fake News Detection: Leveraging Bayesian Optimization and Bidirectional Recurrent Unit. arXiv preprint arXiv: 2502.09097.