Proceedings of Business and Economic Studies, 2026, Volume 9, Issue 1
http://ojs.bbwpublisher.com/index.php/PBES
Online ISSN: 2209-265X
Print ISSN: 2209-2641

Research on Two-Stage Capacity Configuration Optimization of Internet Dispatching Systems: From the Perspective of AI Empowerment and User Behavior

Yunjie Lou*

Baoding University of Technology, Baoding 071000, Hebei, China

**Corresponding author: Yunjie Lou, louyunjiezn@163.com*

Abstract: Against the backdrop of the rapid development of the digital economy, internet dispatching platforms such as food delivery and ride-hailing services have become key urban infrastructure. However, they generally face the core contradiction between dynamic demand fluctuations and rigid service capacity constraints. This paper decomposes the dispatching system into a two-stage closed-loop structure of "waiting and service". Combining queuing theory principles, AI empowerment characteristics, and introducing user loss aversion psychology and reference utility features, a configuration model covering basic capacity and safety capacity is constructed to explore optimal capacity strategies under profit-oriented and welfare-oriented orientations. Numerical examples verify the model's effectiveness. Results show that the optimal capacity consists of basic capacity and safety capacity, with the two-stage safety capacity maintaining a specific matching ratio. Moreover, AI empowerment reduces the basic capacity demand in the waiting stage but requires simultaneous optimization of service stage capacity to avoid new bottlenecks. Consequently, platform positioning and user behavior characteristics significantly affect capacity configuration efficiency. The research conclusions provide theoretical support and practical guidance for dispatching platforms to achieve refined operations and balance efficiency with user experience.

Keywords: Internet dispatching system; Capacity configuration; AI empowerment; User behavior; Queuing theory

Online publication: February 10, 2026

1. Introduction

With the deep integration of digital technology and lifestyle service scenarios, internet dispatching platforms have become the core hub connecting supply and demand sides. The market scale in fields such as food delivery, ride-hailing, and same-city instant services continues to expand. According to iResearch data, China's

instant delivery market transaction volume exceeded 500 billion Chinese Yuan in 2023, and the daily average number of ride-hailing orders exceeded 300 million. The service capacity of dispatching platforms directly affects urban operational efficiency and residents' quality of life.

However, such platforms have always faced prominent operational pain points, For instance, the demand side exhibits significant time-period fluctuations (e.g., morning peak, dinner time), weather sensitivity (e.g., rainfall, high temperature), and sudden scenario characteristics (e.g., large-scale events, holidays). In contrast, the supply-side service capacity (service personnel, equipment resources) has rigid constraints, leading to frequent supply-demand imbalances, such as order backlogs and excessively long user waiting times during peak hours, and service personnel idleness and high operational costs during off-peak hours.

Based on this, this paper divides the internet dispatching system into two stages, "waiting-service", constructs a theoretical model of capacity configuration by combining AI technology empowerment characteristics and user loss aversion psychology, clarifies the optimal ratio of basic capacity to safety capacity, and reveals the law of coordinated matching of two-stage capacity. It provides a new solution for platforms to improve operational efficiency and optimize user experience, while supplementing empirical support for theoretical research in related fields.

2. System model and problem description

2.1. System structure

The service process of an internet dispatching system can be clearly divided into a coupled two-stage closed-loop structure as follows:

(1) Waiting stage: From the moment a user places an order on the platform to when the order enters the system, waits to be assigned to a suitable service personnel (e.g., rider, driver), and is successfully dispatched. The core of this stage is the "queuing" and "matching" of orders;

(2) Service stage: From the moment service personnel accept the order to completing pickup, delivery or service, and finally ending the order. The core of this stage is the "execution" and "completion" of services, which directly determines the actual throughput capacity of the system.

There is a significant interactive effect between the two stages, where the matching quality in the waiting stage affects the route efficiency in the service stage, while the capacity redundancy in the service stage determines the order backlog threshold in the waiting stage, forming a dynamic cycle of "matching-execution-feedback".

2.2. Core problems

The core goal of capacity configuration is to achieve optimal overall system efficiency through resource investment and rule design. Specifically, it is necessary to solve three types of key problems as outlined:

(1) Proportion division between basic capacity and safety capacity: How to set the basic capacity to meet daily demand and the safety capacity to cope with peak fluctuations, balancing operational costs and service stability;

(2) Coordinated matching of two-stage capacity: How to avoid system bottlenecks caused by excess or insufficient capacity in a single stage, and achieve dynamic adaptation between matching efficiency in the waiting stage and execution efficiency in the service stage;

(3) Capacity reconstruction under AI empowerment: How does the efficiency improvement of AI technology in a single stage (e.g., intelligent matching, route optimization) affect the overall capacity configuration, and how to avoid new contradictions arising from technological empowerment.

3. Theoretical model construction

3.1. Basic assumptions

Based on queuing theory principles and the actual operational characteristics of dispatching systems, the following core assumptions are made:

(1) Order arrival distribution: The effective order arrival process follows a Poisson distribution with an arrival rate λ, which is consistent with the order generation rules in scenarios such as food delivery and ride-hailing;

(2) Service rate distribution: The matching rate μ_1 in the waiting stage and the execution rate μ_2 in the service stage both follow exponential distributions, satisfying the memoryless property;

(3) System steady-state assumption: The service rates of both stages are greater than the order arrival rate ($\mu_1 > \lambda$, $\mu_2 > \lambda$) to avoid infinite order backlogs;

(4) AI empowerment assumption: AI technology acts on the waiting stage, increasing the matching rate to $(1 + kAI)\ \mu_1$ ($kAI \geq 0$ is the empowerment level), while generating additional costs CAI positively correlated with the empowerment level.

3.2. Parameter definition

The definition of parameters is as listed:

(1) Demand-side parameters: λ is the effective order arrival rate; h is the user's unit time waiting cost; V is the upper limit of the user's reservation value; p is the service price;

(2) Supply-side parameters: μ_1 is the initial matching rate in the waiting stage; μ_2 is the execution rate in the service stage; c_1 and c_2 are the unit capacity maintenance costs of the two stages, respectively;

(3) AI empowerment parameters: kAI is the AI empowerment level; CAI is the additional cost of AI application;

(4) User behavior parameters: r is the user expectation coefficient; η is the user reference sensitivity, reflecting loss aversion psychology (the degree of aversion to overtime is higher than the satisfaction with early completion).

3.3. Objective function construction

3.3.1. User utility function

User utility is a core indicator of user experience, determined by the reservation value of the service, payment price, waiting cost, and the deviation between actual waiting time and expected waiting time. Considering users' loss aversion characteristics and reference utility features for waiting time, this paper constructs the following user utility function:

$$T(W_1,W_2)=V-p-h(W_1+W_2)+\eta\left(\frac{r}{\mu_1-\lambda}-\frac{1}{\mu_2-\lambda}\right)$$

Where W_1 is the actual time consumed in the waiting stage, W_2 is the actual time consumed in the service stage, and the total waiting time $W = W_1 + W_2$

According to the M/M/1 model of queuing theory, the average time consumed in the waiting stage under steady state is $W_1=1/(\mu_1-\lambda)$, and the average time consumed in the service stage is $W_2=1/(\mu_2-\lambda)$. This utility function includes four core parts: the first part V-p is the user's basic utility, i.e., the value obtained by the user from the service minus the paid price; the second part $h(W_1 + W_2)$ is the user's waiting cost, the longer the waiting time, the lower the user utility; the third part is the user's reference utility, reflecting the impact of the deviation between actual waiting time and expected waiting time on user utility. When the actual total waiting time is less than or equal to the expected waiting time,the user obtains positive reference utility; when the actual total waiting time is greater than the expected waiting time,the user obtains negative reference utility,and the absolute value of the negative utility is larger, reflecting the user's loss aversion psychology.

3.3.2. Platform objective function

Internet dispatching platforms have diverse operational objectives, and different platform positions and market strategies lead to different goal orientations. This paper mainly considers two typical platform objectives: profit orientation and welfare orientation, and constructs corresponding objective functions respectively as follows:

(1) Profit-oriented platforms: Taking profit maximization as the core objective, their profit comes from order revenue. Costs include waiting stage capacity maintenance costs, service stage capacity maintenance costs, and additional AI technology application costs. The profit function is:

$$\Pi=p\lambda-(c_1\mu_1+c_2\mu_2+CAIkAI)$$

Under profit orientation, the platform's capacity configuration strategy will give priority to balancing costs and benefits, and achieve profit maximization by reasonably adjusting the two-stage capacity and AI empowerment level.

(2) Welfare-oriented platforms: Taking social welfare maximization as the core objective, social welfare includes total user utility and platform net profit. The welfare function is:

$$SW=T(W_1,W_2)\lambda+p\lambda-(c_1\mu_1+c_2\mu_2+CAIkAI)$$

Welfare-oriented platforms pay more attention to user experience and overall system efficiency. Their capacity configuration strategy will seek a balance between user utility and platform costs, appropriately sacrificing part of the profit to improve user experience and social welfare.

3.4. Derivation of optimal capacity configuration

To solve the optimal capacity configuration strategy under different goal orientations, this paper uses the Lagrange multiplier method to solve the extreme value of the objective function. In the solving process, the waiting time in the user utility function is replaced by the average time under the steady state of queuing theory, and the matching rate of the waiting stage after AI empowerment is substituted into the objective function to construct the Lagrange function. The partial derivatives of μ_1 and μ_2 are calculated and set to zero to obtain the first-order optimal condition, and then the two-stage optimal capacity configuration formula is derived. Through solving, this paper obtains the following core conclusions.

3.4.1. Law of capacity structure division

Regardless of whether the platform is profit-oriented or welfare-oriented, its optimal capacity consists of two parts: basic capacity and safety capacity. The basic capacity is used to meet the order arrival demand under

steady state, and its size is equal to the effective order arrival rate λ; the safety capacity is used to cope with demand fluctuations, and its size is related to factors such as user waiting cost, user reference sensitivity, service price, and capacity maintenance cost. The optimal capacity configuration formulas for the two stages are:

$$\mu_1^* = \lambda^* + \sqrt{\frac{\lambda(h-\eta r)(p-c_1-c_2)}{vc_1}};$$

$$\mu_2^* = \lambda^* + \sqrt{\frac{\lambda(h+\eta)(p-c_1-c_2)}{\bar{v}c_2}}$$

Where the part inside the square root is the safety capacity, and μ_1^, μ_2^ are the optimal capacities of the two stages, respectively. The optimal matching ratio of the two-stage safety capacity is:

$$\mu_1^s : \mu_2^s = \sqrt{\frac{h-\eta r}{c_1}} : \sqrt{\frac{h+\eta}{c_2}}$$

After AI empowerment,the basic capacity of the waiting stage decreases to λ/(1 + kAI),and the safety capacity matching ratio is adjusted to:

$$\mu_1^{s'} : \mu_2^{s'} = \sqrt{\frac{h-\eta r}{c_1(1+k_{AI})}} : \sqrt{\frac{h+\eta}{c_2}}$$

This conclusion indicates that AI empowerment not only improves the efficiency of the waiting stage but also changes the matching ratio of the two-stage safety capacity. If the platform only improves the AI empowerment level of the waiting stage without synchronously optimizing the capacity of the service stage, the service stage will become a new system bottleneck. Therefore, when applying AI technology, the platform must synchronously adjust the two-stage capacity ratio to achieve the coordinated advancement of technological empowerment and capacity optimization.

4. Key strategies for two-stage capacity configuration

4.1. Capacity configuration strategy for the waiting stage

4.1.1. Dynamic demand response mechanism

The order arrival rate is predicted using historical data and real-time contextual features, including weather conditions, time periods, and special events, and is used to determine dynamic system thresholds. When the number of pending orders exceeds the threshold, a coordinated control strategy is triggered: price incentives are adjusted (e.g., increasing delivery fees), cross-regional scheduling is activated to dispatch nearby idle personnel, and demand guidance is provided by informing customers of expected waiting times. Together, these measures rapidly rebalance supply and demand.

4.1.2. AI-empowered intelligent matching optimization

Order characteristics, including pick-up and delivery locations and time-limit requirements, and service personnel characteristics, such as current location and historical efficiency, are integrated through machine learning algorithms to achieve globally optimal matching. This approach reduces empty driving rates while

enabling differentiated matching priorities for urgent and member orders, thereby improving service quality and the satisfaction of high-value users.

4.1.3. Elastic combination of basic-safety capacity

A fixed base capacity is configured to meet routine daily demand, while a hybrid "full-time + crowdsourcing" model is adopted to provide flexible safety capacity. Crowdsourcing personnel are activated to supplement capacity during peak periods and scaled down during off-peak periods to control costs, thereby maintaining optimal capacity utilization.

4.2. Capacity configuration strategy for the service stage

4.2.1. Route optimization and regional scheduling

Real-time route planning algorithms are used to optimize "one-pickup and multiple-delivery" routes, thereby reducing service time per order. Demand hotspots are predicted using heat-map analysis, enabling advance personnel positioning and a shift from passive order acceptance to active standby, which improves the service rate μ_2.

4.2.2. Service personnel capacity management

Standardized service processes and systematic training programs are established to enhance operational proficiency. Subjective initiative is encouraged through incentive mechanisms, such as efficiency scores and positive review rates. Following AI-enabled efficiency improvements in the waiting stage, it is necessary to synchronously expand safety capacity in the service stage to absorb the increased service pressure generated by higher throughput.

4.2.3. Balance control of cost and efficiency

Safety capacity is adjusted based on the trade-off between unit capacity cost c_2 and user waiting cost h. When h is high, indicating strong user sensitivity to waiting, safety capacity is appropriately increased. Conversely, when c_2 is high, a portion of human capacity is substituted with technical optimization, such as intelligent route planning, to reduce overall costs.

4.3. Two-stage coordinated optimization strategy

A linked monitoring system is established for key indicators across the two stages, including waiting time, order acceptance rate, delivery time limits, and personnel idle rates, to prevent system imbalance arising from single-stage optimization. For example, when AI-driven matching efficiency in the waiting stage is improved, service-stage capacity must be synchronously expanded to eliminate the emerging bottleneck of "fast matching but slow execution." Capacity configuration ratios are then dynamically adjusted through real-time data feedback to achieve global system optimization.

5. Numerical examples and management implications

5.1. Numerical example verification

Parameters are set based on the actual operational data of the dispatching system: $h = 1.6$, $\alpha = 0.6$, $r = 1.1$, $c_1 = 0.3$, $c_2 = 0.5$, $V = 6$, $\lambda = 10$, $CAI = 1$. The following key conclusions are obtained through numerical simulation:

(1) When the AI empowerment level kAI increases to 1, the basic capacity of the waiting stage decreases by more than 30%, and the safety capacity is saved by 25%, but the safety capacity of the service stage needs to be increased by 15% to match the efficiency;
(2) Profit-oriented platforms are more inclined to control safety capacity investment, while welfare-oriented platforms will configure more safety capacity to reduce user waiting time;
(3) When the user expectation coefficient r is in the interval [0.8,1.3], it is difficult for the platform to accurately judge user sensitivity, which is prone to capacity configuration deviations, and strategies need to be adjusted through dynamic monitoring.

5.2. Management implications

AI empowerment needs to consider the whole-process optimization. Only improving the matching efficiency of the waiting stage may trigger bottlenecks in the service stage. The platform should synchronously adjust the two-stage capacity ratio according to the kAI value to avoid "one-sided optimization" of technological empowerment. Capacity configuration should be in line with user psychology. Based on users' reference sensitivity and loss aversion characteristics for waiting time, prioritize configuring safety capacity in the service stage to reduce the risk of waiting overtime. Profit-oriented platforms can dynamically adjust safety capacity according to the cost-benefit ratio; welfare-oriented platforms need to seek a balance between capacity investment and user experience, and appropriately increase the proportion of basic capacity.

6. Conclusion

This paper divides the internet dispatching system into two stages, "waiting-service", constructs a theoretical model of capacity configuration by combining queuing theory, AI empowerment characteristics, and user behavior psychology, and reveals the optimal configuration law of basic capacity and safety capacity and the two-stage coordination mechanism. Research shows that the optimal capacity configuration needs to comprehensively consider the platform's operational objectives, user behavior characteristics, and technological empowerment level, and achieve supply-demand balance and efficiency improvement through strategies such as dynamic demand response, AI intelligent matching, route optimization, and coordinated scheduling. Future research can further expand the scenario boundaries, incorporate factors such as multi-regional coordinated scheduling and service personnel heterogeneity (efficiency differences, preference differences), and construct a multi-objective capacity configuration model closer to reality. Meanwhile, conduct A/B testing combined with empirical data to verify the applicability of the model in different dispatching scenarios, providing more accurate decision support for the refined operation of platforms.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Hong W, 2024, Research on the Business Model of Steel Business of Fujian SG Company Based on the IoT Cloud Commerce Platform, 32–38.

[2] Zhou Y, Li F, 2022, Problems and Challenges Faced by New Retail Operation Management. Journal of Systems & Management, 31(5): 12–13.

[3] Liu S, 2023, Optimization of Unmanned Vehicle Delivery Operation Strategy for Online Retailers, thesis, Beijing Jiaotong University, 78–89.

[4] Gu Y, 2024, Research on the Dispatching Mode of Residential Decoration Trades Based on BIM+AR, China Architecture & Building Press, Beijing, 45–68.

[5] Yang J, 2021, Design and Implementation of Fresh E-commerce Order Fulfillment System, Publishing House of Electronics Industry, Beijing, 20–21.

[6] Yao Y, Xia B, 2014, Application of Phase Frequency Feature Group Delay Algorithm in Database Differential Access. Computer Simulation, 31(12): 238–241.

[7] Gamelin F, Baquet G, Berthoin S, et al., 2009, Effect of Low-Intensity Training on the Triacylglycerol Oxidation Rate During Exercise. Journal of Applied Physiology, 2009(105): 731–738.

[8] Jackson D, Firtko A, Edenborough M, 2009, Personal Resilience as a Strategy for Surviving and Thriving in the Face of Workplace Adversity: A Literature Review. Journal of Advanced Nursing, 60(1): 1–9.

[9] Hargreve M, Jensen A, Nielsen T, et al., 2015, Maternal Use of Fertility Drugs and Risk of Childhood Acute Lymphoblastic Leukemia. International Journal of Cancer, 136(8): 1931–1939.

[10] Schneider Z, Whitehead D, Elliott D, 2009, Nursing and Midwifery Research: Methods and Appraisal for Evidence-Based Practice, 3rd edn, Elsevier Australia, Marrickville, NSW.

[11] Davis M, Charles L, Curry M, et al., 2013, Challenging Spatial Norms, Routledge, London.

[12] Knowles M, 1986, Independent Study, In Using Learning Contracts, Jossey-Bass, San Francisco: 89–96.

[13] Zhang S, Liaw L, Ruppenicker J, 1999, Proceedings of the Twenty-Fifth Annual Meeting of the Berkeley Linguistics Society, February 12-15, 1999: General Session and Parasession on Loan Word Phenomena. 2000, Berkeley Linguistics Society, Berkeley.

[14] Bukowski R, 2008, Prognostic Factors for Survival in Metastatic Renal Cell Carcinoma: Update 2008. Innovations and Challenges in Renal Cancer: Proceedings of the Third Cambridge Conference. Cancer, 115(10): 2273.

[15] Este J, Warren C, Connor L, et al., 2009, Life in the Clickstream: The Future of Journalism. Media, Entertainment and Arts Alliance, 2009.

[16] Developing an Argument, 2025, https://writing.princeton.edu/resources/academic-writing/constructing-arguments.

[17] Gale L, 2000, The Relationship Between Leadership and Employee Empowerment for Successful Total Quality Management, thesis, University of Western Sydney.