



PROCEEDINGS

IPIC 2020

**7th International Physical Internet
Conference**

**HYPERCONNECTING THE WORLD WITH
PHYSICAL INTERNET**



Towards a smart hyperconnected era of efficient and sustainable logistics, supply chains and transportation

Organizing Committee

Organizing Committee Chair

Hao LUO: Shenzhen University

Organizing Committee Co-Chair

Yi ZHAO: Shenzhen University

Ray ZHONG: The University of Hong Kong

Student Helper

Xuan YANG: The University of Hong Kong

Qunye ZHANG: Shenzhen University

Xuejiao WANG: Shenzhen University

Yi XIA: Shenzhen University

Sangsang WU: Shenzhen University

Mingjie GONG: Shenzhen University

Yuting WANG: Shenzhen University

Jianxin XIAO: Shenzhen University

Luyao CHENG: Shenzhen University

Chao WANG: Shenzhen University

Huan LIU: Shenzhen University

Huanguang LIAO: Shenzhen University

Mengli WANG: Shenzhen University

Xubin XIN: Shenzhen University

Jie MEI: Shenzhen University

Jianbo HE: Shenzhen University

Bowen ZHANG: Shenzhen University

Jueying XIANG: Shenzhen University

Siyu TIAN: Shenzhen University

Qingyi ZHANG: Shenzhen University

Nan LIU: Shenzhen University

Yujie WANG: Shenzhen University

Scientific Committee

Scientific Committee Chair

Benoit Montreuil: Georgia Institute of Technology

Scientific Committee Co-Chair (International)

Shenle PAN: MINES Paris Tech

Scientific Committee Co-Chair (Local)

Lijun MA: Shenzhen University

Scientific Committee Members

Yaping ZHAO: Shenzhen University

Zelong YI: Shenzhen University

Yelin FU: Shenzhen University

Special Session Chair

Wei QIN: Shanghai Jiao Tong University

Ting QU: Jinan University

Pengyu YAN: University of Electronic Science and Technology of China

Suxiu XU: Jinan University

Shenle PAN: MINES Paris Tech

Yue ZHAI: Beijing Jiao Tong University

Hing Kai Chan: University of Nottingham Ningbo, China

Zelong YI: Shenzhen University

Kin Keung Lai: Shenzhen University

Ray ZHANG: The University of Hong Kong

Clinton Liu: MCG Canada

Abraham Zhang: University of Essex

Xiao LIN: Department of Transport & Planning, Delft University of Technology

Industrial Committee

Xiang T.R. KONG: Shenzhen University

Fernando Liesa: ALICE, Alliance for Logistics Innovation through Collaboration in Europe

David LEUNG: Logistics and Supply Chain MultiTech R&D Centre

Contents

Decentralized and centralized transport and logistics carbon emission optimization and emission norms for the transport and logistics sector	1
<i>Igor Davydenko, Meike Hopman and Jordy Spreen</i>	
Design and Evaluation of Routing Artifacts as a Part of the Physical Internet Framework	14
<i>Steffen Kaup, André Ludwig and Bogdan Franczyk</i>	
Complexity of rules in crowdsourced deliveries and its level of intrusiveness on participants: An experimental case study in the Netherlands.....	30
<i>Xiao Lin, Yoshinari Nishiki and Lóránt A. Tavasszy</i>	
Hierarchical Staffing Problem in Nursing Homes.....	41
<i>Ting Zhang, Shuqing Liu, Ping Feng, Yali Zheng and Wenge Chen</i>	
Resource efficiency optimization-oriented digital twin unmanned warehouse system.....	52
<i>Peihan Wen, Xuqian Ye and Yiyang Liu</i>	
Application of Internet of Things into smart home scheduling.....	66
<i>Yun Huang and Fan Gao</i>	
Integrated Production and Maintenance Scheduling using Memetic algorithm under Time-of-use Electricity tariffs.....	73
<i>Tong Ning and Jian Chen</i>	
Physical Internet-enabled synchronized optimization for Milk-run transportation and Cross-docking warehouse in industrial park.....	86
<i>Yuanxin Lin, Ting Qu, Kai Zhang and George Q Huang</i>	
Multi-agent reinforcement learning-based dynamic task assignment for vehicles in urban transportation physical internet.....	101
<i>Wei Qin, Yanning Sun, Zilong Zhuang, Zhiyao Lu and Yaoming Zhou</i>	
A Two-Stage Production Planning Model for Perishable Products Under Uncertainty.....	116
<i>Kin Keung Lai and Ming Wang</i>	

Dynamic Optimal Approach for an Electric Taxi Fleet's Charging and Order-service Schemes.....	126
<i>Kaize Yu, Pengyu Yan and Zhibin Chen</i>	
Analysis of a Physical Internet enabled parking slot management system.....	141
<i>Bingqing Tan, Suxiu Xu and Kai Kang</i>	
Synchroperation in Industry 4.0 Manufacturing.....	152
<i>Daqiang Guo, Mingxing Li, Zhongyuan Lyu, Kai Kang, Wei Wu, Ray Y. Zhong and George Q. Huang</i>	
Graduation Intelligent Manufacturing System for Advanced Planning and Scheduling in PI-enabled Hyperconnected Fixed- Position Assembly Islands.....	185
<i>Mingxing Li, Daqiang Guo, Ray Zhong and G.Q. Huang</i>	
Design and decision optimization of robot shuttle system.....	204
<i>Wei Wang, Yaohua Wu and Ming Li</i>	
Data-driven analytics-based capacity management for hyperconnected third-party logistics providers.....	222
<i>Jana Boerger and Benoit Montreuil</i>	

Decentralized and centralized transport and logistics carbon emission optimization and emission norms for the transport and logistics sector

I.Y. Davydenko¹, W.M.M. Hopman, J.S. Spreen

TNO Sustainable Transport and Logistics, the Hague, the Netherlands

¹Corresponding author, igor.davydenko@tno.nl

Abstract: Transport and logistics is one of the most important economic sectors contributing to the climate change. By the nature of transport and logistics operations, the sector is one of the most difficult ones to decarbonize. This paper proposes using carbon footprinting tools to optimize logistics operations with respect to emissions, and to setup government-led emission norms for the transport and logistics sector. Carbon footprinting can be used for operational decision making, such as those envisioned by the concept of physical internet, as well as in the classical operations research centralized optimization. The paper shows conceptually how carbon footprinting indicators are applicable for the traditional logistics optimization and for the decentralized optimization of operations. The governments can further speed up the process by setting emission norms for the transport and logistics. This paper shows that the carbon footprinting methods provide sufficient input for both logistics optimization and the norms. The carbon footprinting indicators are discussed and incorporated into the mathematical formulations of logistics optimization; the same carbon footprinting data is used for the setup of carbon emission norms in logistics.

Keywords: carbon footprint, carbon footprint minimization, emission monitoring, emission norms, transport and logistics optimization, centralized optimization, decentralized optimization

1. Introduction

A longer-term challenge of decarbonization of transport and logistics is huge. The climate change movement progresses from an acknowledgement of the problem to undertaking of actions. Depending on the ambition, decarbonization actions can be set to reduce emissions by 60% in 2050 compared with the baseline of 1990, effectively meaning a factor 6 increase in carbon productivity of the system (Smokers et al., 2019). A larger ambition can be set if 95% of emissions are to be reduced by 2050, which in essence means complete decarbonization of the system, or simply said, factor infinity. This means that for the long-term the transport and logistics sector has to be reorganized based on zero-emission technology.

A medium term goal of the European Commission is to reduce Greenhouse Gas (GHG) emissions by at least 40% in 2030 compared to the 1990 levels, as provided in the EU 2030 climate & energy framework. On a national level, the Netherlands the mobility sector is to achieve a 22% emission reduction by 2030 compared to a no action business as usual scenario (Klimaatakkoord, 2019; Hekkenberg and Koelemeijer 2018). Both the EU and national action plans confirm that for the medium term (action 2030), a complete decarbonization seems to be unfeasible due to a number of reasons, such as technological immaturity of zero emission vehicles, market unavailability of zero emission vehicles, insufficiently decarbonized generation of electricity, and lack of infrastructure. There are also some transport areas, such as aviation and long distance transport, which are hard to electrify. These considerations mean that in medium term a mix of carbon intensive and zero emission solutions will coexist.

Practically, decarbonization of transport and logistics operations can be facilitated by two forces: private and public parties. The first force comes from decision makers (e.g. planners and supporting software) working on behalf of private or corporate parties. These parties pursue the goal of logistics operations optimization within some certain boundary conditions. At this moment, the paradigm of operationalization of decision making in logistics, such as Synchronomodality (e.g. Tavasszy et al., 2017; van Rissen et al., 2015) and physical internet (Montreuil, 2011) gain especial attention due to increased efforts on decarbonization of logistics operations.

The second force comes from the governmental bodies (public parties), who can influence the system by the fiscal means (e.g. fuel taxes, vehicle taxes), as well as by the means of permits and norms. For both private and public types of decision makers there is a need for objective information on GHG emissions, with a difference in the aggregation level: the private decision makers will mainly need more disaggregated and specific data, while the public decision makers will mainly need more system-wide aggregated data.

In this paper we do not consider cases in which all the needed information is available to the decision maker, as for instance, may be the case within a transport company. In that case the decision maker can, for instance, make sure that an optimum route, within business constraints, is driven by the fleet. We concentrate on the state-of-the-art where the GHG emission performance of third parties is not directly known to the users of the services and where there is no good aggregated information on the sector-specific GHG emission performance of constituting companies.

This paper is structured as follows. Chapter 2 provides mathematical formulations on how to include GHG emissions into logistics optimization decisions. These formulations are applicable at the level of decentralized (and possibly distributed decision making) as being considered in the context of physical internet, as well as at the traditional level of centralized logistics optimization, as a well-established part of operations research. Chapter 3 provides formulations for indicators that can form a basis for the government-regulated norms for logistics emissions. The chapter further discusses the ways on how the norms can be formulated in practice. Chapter 4 provides ideas on data and governance infrastructure that need to be put in place to collect relevant data for both logistics optimization and setup of norms. Chapter 5 proposes a data collection and processing to service both logistics optimization and policy making purposes. Chapter 6 provides conclusions and outlines directions for further research to close still existing gaps in methodologies and knowledge.

2. Decentralized and centralized carbon optimization of transport and logistics by private parties

Optimization of transport and logistics is well studied and an integral part of Operations Research. For instance, a widely cited review of the literature on facility location and supply chain management (Melo et al., 2009), contains 139 references to the peer-reviewed works on this problem. The logistics and supply chain optimization traditionally balances two conflicting goals: provision of the clients with the desired service level, while minimizing expenses and costs associated with the logistics operations (Davydenko, 2015). The classical basic tradeoffs involved in logistics optimization are the balance between transport and stock keeping costs (e.g. economic order quantity, Blumenfeld et al., 1985, Goyal, 1985); the balance between the speed and cost of services (e.g. Tavasszy et al., 2011). In a broader sense, there is a tradeoff between the cost of sourcing products versus transport costs from the production locations to the consumption locations (e.g. Moses, 1958), as it can be more attractive to source products cheaply overseas and pay more for the transport services.

The logistics decisions lay mostly in the realm of private or corporate decision makers. Depending on the decision to be made, the choice set can be relatively small (e.g. the choice on transport mode to be used to transport goods) or the choice set can be relatively large (e.g. the choice on supply chain organization). The last one often involves solving a facility location problem.

This paper introduces explicit inclusions of GHG emissions into the optimization of logistics operation by private parties. The GHG emissions can be assigned a certain monetary value, proportionally to the volume of GHG emitted and the cost of one ton of the CO₂ or CO_{2eq} emissions. The inclusion of CO₂ costs into the decision process can be done at both operational and strategic levels. At the operational level, a set of transport options can be created, for example a set of possible ways to transport containers from the port using a direct road connection or using intermodal transport, involving inland navigation or train line haul with a subsequent last mile road leg from the intermodal terminal to the final destination. Another example is the choice that a parcel “can make” with respect to the vehicle in which the parcel will travel to the end destination. Equation (1) provides a simple formulation for the disutility function for a choice set that includes transport, time and emission related costs.

$$U_i = C_i + T_i * VOT + C_{CO_2e} * W_i, i = 1..n \quad (1)$$

Where:

- U_i : total disutility of option i in €/unit (e.g. ton, m³, container, parcel, ...)
- C_i : out-of-pocket cost of option i in €/unit paid to the service provider(s)
- T_i : time it takes per transported unit to perform operations related to option i
- VOT : value of time in € per time unit in accordance to T_i
- C_{CO_2eq} : cost of a ton of CO₂ or CO_{2eq} emissions
- W_i : total weight (ton) of CO₂ or CO_{2eq} emissions per transported unit related to option i

The cost C_{CO_2eq} may be a fictive cost related to a company's internal accountancy. The total number of options is expressed as n . The decision maker chooses the option for which the U_i value is the smallest. For the purpose of illustration, the disutility function is kept to simplicity.

Equation (1) expresses the way on how to include the costs of CO₂ emissions into the operational environments. With the rise of self-organization and the concept of the physical internet, it is important to equip distributed decision makers with information about GHG emissions related to the choice set that these decision makers are to explore in the process of taking decisions. Equation (1) is also an example on how to incorporate the true costs of GHG emissions into the operational logic of distributed decision makers, which is a cornerstone of the concept of self-organizing logistics and the physical internet. This formulation is also suitable for incorporation into transport, using for instance, multinomial logit discrete choice model formulations (e.g. Bhat, 2000).

Similarly to the operational decisions, the true costs of GHG emissions can be included at strategic level, for example when long term decisions are made on the location of facilities, such as warehouses, distribution centers, crossdocks, and other facilities. In line with the classical facility location formulations (Campbel, 1994; and Haug, 1985), this can be formulated as an integer programming problem as shown in equations (2) and (3). The choice set includes n possible locations where a facility can be placed with the goal of systemic total cost optimization. The cost function can include transport costs, facility costs and emission costs, as presented in equation (2), but other, more broad formulations are also possible.

$$\min \left(F_i + C_i^{in} + C_i^{out} + C_{CO_{2eq}} * W_i^{in} + C_{CO_{2eq}} * W_i^{out} \right) V_i * z_i, i = 1..n \quad (2)$$

$$\text{s.t. } \sum_{i=1}^n V_i * z_i = V \quad (3)$$

Where:

- F_i : cost of facility i in € per volume of freight
- C_i^{in} : inbound out-of-pocket transport cost for facility i in € per volume of freight
- C_i^{out} : outbound out-of-pocket transport cost for facility i in € per volume of freight
- $C_{CO_{2eq}}$: cost of a ton of CO₂ or CO_{2eq} emissions
- W_i^{in} : inbound weight of CO₂ or CO_{2eq} emissions per volume of freight for facility i
- W_i^{out} : outbound weight of CO₂ or CO_{2eq} emissions per volume of freight for facility i
- V_i : volume of freight (annually) flowing through facility i
- z_i : binary variable ($z_i = 0,1$) indicating whether facility i should be built
- V : total volume that should be shipped through the system

This formulation ((2) and (3)) can be extended with other cost components and service requirements such as, for example, stock keeping costs and speed of service. Similarly to the disutility formulation (1), the integer program is kept to simplicity for the purpose of illustration.

In both operational (equation (1)) and strategic (equation (2) and (3)) cases, the amount of CO_{2eq} emitted (W_i, W_i^{in}, W_i^{out}) is not yet known as the transport operations will take place in the future. The ex-ante amount of CO_{2eq} to be emitted can be estimated in the following three ways:

- 1) Using assumptions about the organization of transport operations;
- 2) Using default data, such as industry average CO₂ emissions per ton-kilometer transported;
- 3) Using service provider specific emission factors based on the ex-post data of the service provider in question.

Emission estimation in accordance to option 1) is the least feasible among the three options. To provide better estimations for a specific organization than the industry average assessments of GHG emissions (option 2)), some knowledge and data on the organization of operations will be required. Moreover, there may be involved some computationally challenging tasks, such as determining the route and possibly solving a traveling salesman problem for each option. Such an approach is not feasible to be included into the integer program (equations (2) and (3)), nor is it reasonable to assume that distributed decision makers, as specified in equation (1), are capable of gathering the data and performing these computations. Option 2) is the easiest to apply, but has a drawback that it does not include any data on performance of specific service providers, nor can it take into account any local specifics. Option 3) allows using measured ex-post data for determining the future course of action – there is no guarantee that performance will be the same as measured in the previous period, but it is the best available approximation on a set of limited information for the future performance. Moreover, option 3) allows distinguishing between different service providers allowing to informatively choose the best performing one.

Private parties need information on GHG emissions for both operational (equation (1)) and strategic (equation (2) and (3)) decisions. Option 3) is the most suitable way to estimate the ex-ante amount of CO_{2eq} to be emitted. An additional advantage of option 3) is that this

option can also be used as data input towards the formulation of GHG emission norms by public parties, as discussed in the following chapter.

3. Considerations on formulation of GHG emission norms by authorities

Decarbonization of transport and logistics is facilitated by two types of forces: private parties (as discussed in Chapter 2) and public forces (as discussed in this chapter). At the policy level, the question of regulation of transport and logistics emissions has gotten a new impetus. Similarly to the regulation of vehicle emissions, there is an ongoing discussion on an introduction of emission norms for the transport and logistics sector. Additionally to the political challenges, the policymakers face the technical challenge on how to set up a norming scheme. Specifically, what has to be the basis of a norm, i.e. what to measure, in what units and how? Once these questions have been answered, the policymakers will face the challenge of getting the baseline data right. Specifically, how to get adequate information about the current state of the industry with respect to quantitative data on the chosen measure? How to segment diverse logistics sectors into homogeneous segments where a norm can be applied?

Logistics performance with respect to GHG emissions can be measured as the amount of CO_{2eq} emitted per unit of transport activity. Different indicators exist that are aimed at different types of stakeholders, however, two large classes of the indicators can be distinguished (Davydenko et al., 2019).

1. Carbon efficiency of a service provider: gCO_{2eq} per unit of freight per unit of distance, for instance gCO_{2eq} per ton-kilometre or gCO_{2eq} per m³-kilometre of transport carried.
2. Carbon efficiency of a shipper: gCO_{2eq} per unit of freight, for instance gCO_{2eq} per ton or gCO_{2eq} per m³ shipped.

Specifications of the unit of freight are usually limited to the weight (tonnes), volumes (cubic meters), TEU or containers, pallets and packages, although other units of freight may be used. The most common among them is the weight unit. The unit of distance is kilometer (Imperial unit is mile) and there are different ways to measure the distance, which is discussed in more detail in Chapter 4.

Based on these considerations, there can be two types of norms proposed. The first type of norm is related to the operations of service providers who work within the logistics industry. The service providers' related norm will be expressed in gCO_{2eq} per ton-kilometre transported. The second type of norm is related to the operations of shippers – the users of transport and logistics services. The shippers' indicator will be expressed in gCO_{2eq} per ton shipped. The shippers' indicator combines the service provider's carbon efficiency with the overall organization of the shipper's supply chain. In other words, the less spatially stretched the shipper's supply chain and the more efficient the service provider of their choosing, the better is the shipper's indicator.

The process of setting the norms includes determining the carbon performance of market parties in the segment. A possible approach to setting up the norms is to determine the distribution of the emission values by the companies active in the segment and to set the targets such that the worst performing companies will have to improve or go out of business. Concentration on the worst performing companies has two advantages: first, it removes the worst performing operators (i.e. those that emit disproportionately more gCO_{2eq} per ton-

kilometre or per ton shipped) and second, by removing the worst performing operators from the market, the total emissions will be lowered, as well as the average level of emissions. The process of target setting can be organized in a way that, for instance, performance of worse than two standard deviations over the mean is forbidden, affecting around 5% of the company population, depending on the form of distribution. Once the new norm is set, it can be revised over a period (e.g. one year) in a similar way, thus creating the pressure on continuous improvement in the market, see figure 1 for an illustration.

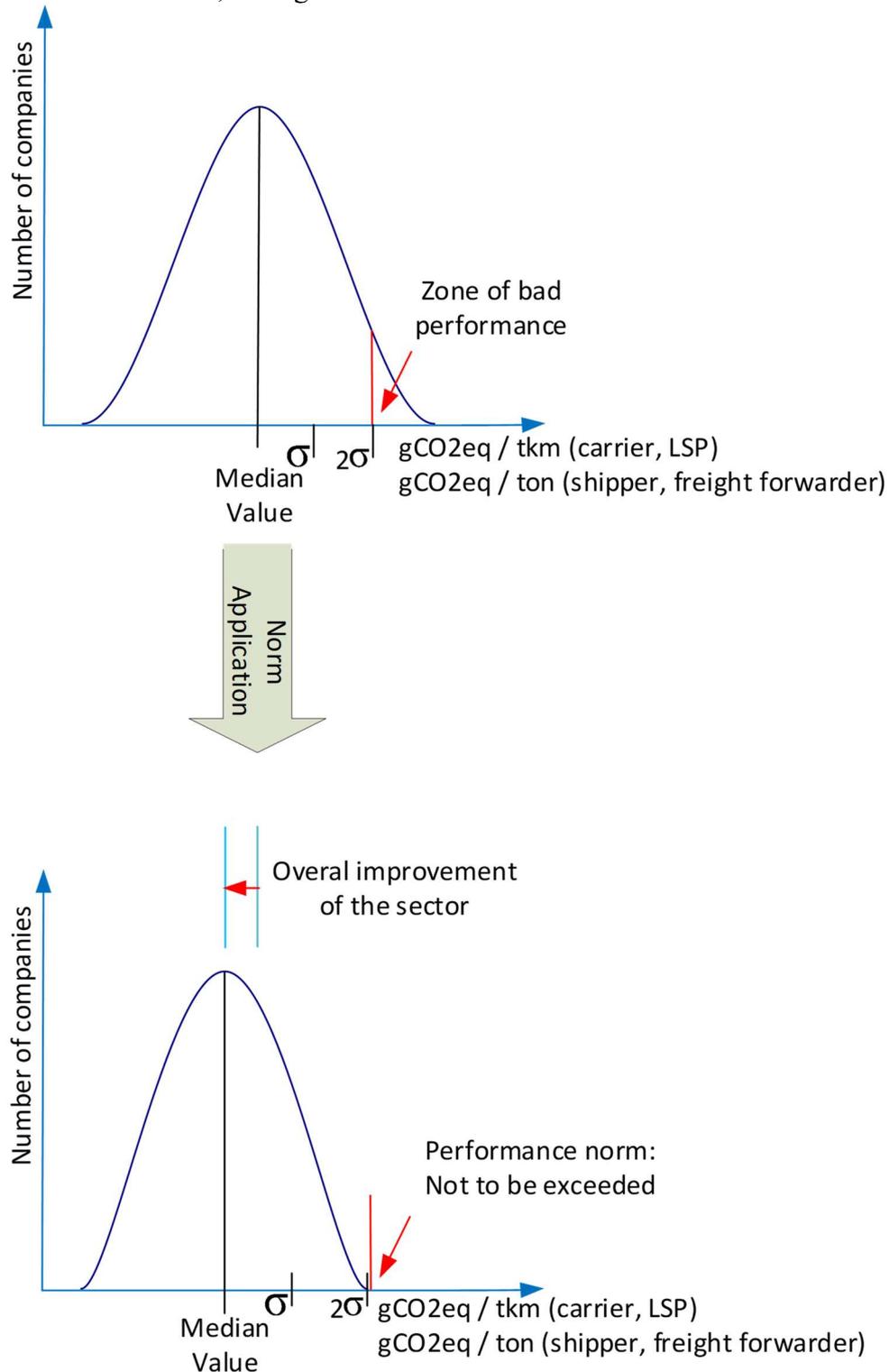


Figure 1. Example of a way for norm set up

This paper is discussing the issue of logistics segmentation that is related to the fact that logistics operations are heterogenous in their nature. The different segments are not directly comparable with each other in terms of CO₂ emissions. For instance, the average fuel consumption per ton-kilometer of goods shipped in a van is 10 times bigger than the same indicator for the goods transported in a 40-ton truck (Greene and Lewis, 2016). Therefore, a proper segmentation of the transport market is a condition for a norming scheme and deserves a dedicated consideration.

4. Data infrastructure for GHG emission optimization and GHG emission norms

As we discussed in chapters 2 and 3, for both logistics process optimization and policy applications, a measure of GHG emissions related to transport activity is the needed input into the decision making process.

4.1. Transport activity

Transport activity is measured in terms of freight units transported over distance units.

Units of distance. Five fundamental distance measures can be distinguished:

- 1) **Great Circle Distance (GCD).** The great circle distance is the shortest distance between two points on the surface of the Earth, measured along the surface of the Earth. It is also known as the “as the crow flies” distance: this distance does not consider any infrastructure, so two points are connected directly, as if there is a straight road between them. The GCD is the most suitable measure for distance for the purpose of carbon footprinting as it looks at the net transport work independent of the chosen modality, infrastructure density and routing of the goods flow. It is the only measure that leads to a correct calculation of the impact of changes in routing or modalities on the carbon footprint. It is also the “easiest” distance measure from an administration and data requirements point of view, as there is no need to keep track of the routes that the vehicles travelled (Davydenko et al, 2019);
- 2) **Actually Driven Distance (ADD).** The actually driven distance is the distance travelled by the vehicle. This distance can be measured by the vehicle’s odometer. The ADD is the most intuitively understandable distance: for this reason it has deep usage roots. For instance, transport statistics is expressed in ton-kilometres actually driven and the companies are used to reporting to the statistics bureaus in this manner. Also, some transport companies charge their clients based on travelled distances (Davydenko et al). The ADD has a number of drawbacks with respect to establishing GHG emission performance indicators. First, the ADD does not reflect on efficiency of the routes, as for instance, unnecessary kilometres are not penalised. The ADD can even encourage more kilometres to be driven in case emissions made while making those kilometres are less than the average emissions. Second, for the logistics optimization purposes, as discussed in Chapter 2, the ADD is not known ex-ante, estimation of this distance requires assumptions and optimization, which are not possible or desirable in the distributed decision environment, nor it is suitable for the integer programming. Third, the ADD has to be logged and stored by the carrier; this distance measure is not generally available to any other party than the carrier. Despite the fact that the ADD is often

used in carbon reporting and accountancy, the abovementioned drawbacks make the ADD distance unit an inferior unit compared to the GCD.

- 3) **Planned Distance (PD).** The planned distance is the distance that a shipment is expected to follow in a vehicle as the route of the vehicle is determined by the planning software. The PD as a distance unit measure for the GHG emission measure indicator is equivalent to the ADD and, thus despite wide use in carbon reporting and accountancy, it is inferior compared to the GCD unit measure.
- 4) **Shortest Feasible Distance (SFD).** The shortest feasible distance is the shortest distance between two places on a mode-specific network. The SFD can be computed by any party having access to the network specifications and software capable of computing shortest path. The SFD is a physical distance over infrastructure, thus more similar to the GCD distance measure. Compared to the GCD, it has three drawbacks: 1) it is mode-dependent, 2) it needs special software to be computed and 3) it changes when the network is adjusted. This makes the use of SFD slightly less attractive than the GCD.
- 5) **Fastest Distance (FD).** The fastest distance is the distance of the route that allows travelling from the departure point to the arrival point at a minimum time. The FD is essentially equal to the SFD, with the only difference that instead of distance, travel time is minimized while determining the FD. The GCD is more preferable unit than the FD due to the same drawbacks as those of the SFD.

Units of freight. Units of freight can be characterized by their physical properties, such as weight and volume, as well as specific industrial conventional load units.

- 1) **Weight (tons).** Weight is the most common unit of freight. Weight is relatively easy to obtain by weighing the goods; if it is not practical to weigh the goods, then the total weight is the sum of weights of individual items.
- 2) **Volume (m³).** The volume of goods is also a common measure of freight, especially in case of volume-limited operations, or freight with a high volume to weight ration. Volume is not as often measured as weight, however, for some operations like parcel deliveries, volume is more common than weight.
- 3) **Load units.** The most used load unit is container, measured in 20-foot container equivalents (TEU) for shipping, and in LD-3 and other containers in aircraft operations. Other load units, such as pallets and individual SKU's or parcels can also be used. The load unit measures are common for arrangement of pallet and shipping container transport.

As we considered different measures to determine transport activity, the measure based on the Great Circle Distance and weight transported can be considered the most useful for logistics optimization and a setup of logistics emission norms. In case other than weight units of freight are universally used across the segment, it can be acceptable to use m³ * distance GCD as the common transport activity measure for that specific segment.

4.2. GHG emission measures

Green House Gas emissions are measured as the weight of CO₂-equivalent emissions made while carrying out certain transport activity. The measured GHG emissions should include all vehicle operations, including empty runs, repositioning and other non-revenue use that is essential for conduction of primary business activities.

In practice, the GHG emission weight is determined by multiplication of volume of fuel burned (or the amount of electricity used) by an emission factor, which specifies the weight of CO₂-equivalents released into the atmosphere by burning one liter or one kilogram of fuel, or by using one kilowatt hour of electricity. A practical way to determine the weight of GHG emissions is to get fuel purchasing data (or charging data if applicable) over a period and to multiply the amount of fuel or electricity used in that period by a relevant emission factor. As in many cases tanking does not happen every day, relatively large rounding error may occur if aggregated over a short period, in many cases it is reasonable to aggregate fuel and electricity use for periods of at least one month. Fuel and electricity use aggregation of one year has an advantage of smoothing out seasonal patterns of energy use and seasonal patterns of transport service demand.

For determination of logistics emission performance indicators, as discussed in Chapter 2 and 3, the emission data has to be normalized per unit of transport work. It is important to ensure that there is a unique and unambiguous match between transport activity carried out and fuel (or electricity) use. In other words, it must be ensured that the vehicles are used only for services falling within the scope of transport activities, and that transport activities are carried out only within the scope of measured fuel or electricity use.

5. Proposal for data collection process and data processing design

The logistics emission calculation tools (e.g. BigMile, EcoTransIT World, EPA's SmartWay, TK'Blue and others) together with the public data collection institutes, such as the Dutch Statistics Bureau CBS, can provide the necessary physical and institutional infrastructure for emission data collection, processing and analysis. At the micro level, where decisions are made on the optimization of operations, and the macro (policy) level, where the emission norms are to be set, the emissions are computed and normalized to the indicators discussed earlier in the paper. The following can be considered as the main data collection requirements.

1. **For micro level decisions**, such as those discussed in Chapter 2, the data collection arrangement should provide an easy to use computation of GHG emissions related to certain logistics choices based on primary data of service providers. This this can be realized by 3rd party platforms that collect micro data from the businesses and which, authorized by the data owners, can share emission data with intended recipients or the public.
2. **For macro (policy) level decisions on emission norms**, such as those discussed in Chapter 3, the data collection process should provide sufficiently aggregated data on the GHG performance of businesses. This can be done through comprehensive survey(s) of transport and logistics.

The basis GHG emission KPI in the transport networks of carriers is gCO_{2eq} per ton-kilometer GCD transported, and for the shippers the basis KPI is gCO_{2eq} per ton of goods shipped, which can be determined in accordance to the discussion in Chapter 4. The carriers

are in principle capable to compute this indicator by themselves and subsequently publish it in a form as, for instance, proposed by the GLEC declaration. In some cases, the carriers are not possessing all the data necessary to compute this KPI (see more details on the absence of cargo weight data by the carriers in LEARN D4.4 (Davydenko et al., 2018) – the analysis of around 30 carbon footprint implementations at industrial companies). In this case tools that help collect data (e.g. electronic bill of lading, aggregation of different data sources, intercompany links) may solve the problem.

Another challenging issue that needs to be overcome is the sensitivity of GHG emission data. From the emission data the amount of fuel used can be determined, and thus fuel costs, which is one of the most important expenses of the service providers. Some of them may not be willing to share this information broadly. Therefore, for the purpose of policy-related data collection, the Statistics Bureaus (e.g. CBS) can be asked to collect emission performance data, in addition to the data they collect on, for example, goods flows. This may use the existing organizational and survey infrastructure with strict data privacy norms.

For the operational and strategic decisions by the users of transport services, the logistics emission calculation tools can be extended towards data services (e.g. SmartWay can be considered one of those, although it does not compute the specific indicators discussed in this paper) that allow communication of emission performance data between market parties. In this way the data owners can restrict and specify the list of other parties who may be provided limited access to their data. For instance, for a specified origin and destination, the service may return a number of options (e.g. modalities and carriers) with the emission data related to each of the choices.

6. Conclusions and outline for further research

This paper has provided a discussion on how to include GHG emissions in logistics decisions and optimization of logistics operations. The optimization of logistics processes can be done locally and operationally, possibly by the distributed decision makers, such as it is foreseen in the concept of physical internet. The optimization can be done in a classical way, globally or centrally, where integer programming can be used for determining an optimum supply and transport chain designs. The added value of this discussion is that, in addition to the usual optimization goal of cost reduction and maximization of the service level, the resulting GHG emissions are taken explicitly into account and directly impact the outcome of operational and strategic decisions. Depending on the costs of a ton of CO₂ emission constant used, the formulations provided in the paper may shift decisions from the cheapest solutions within service constraints to the least polluting ones within the same constraints.

Similarly to the emission norms for vehicles, there is an ongoing discussion on regulating logistics emissions through formulation of emission norms for logistics operations. This paper provides a discussion on how to set up logistics GHG emission norm regulations using carbon footprinting methods developed for the micro level, i.e. bringing carbon footprinting to the macro level, at which policy makers work.

To realize both logistics optimization and to set up norms for GHG emissions in logistics operations, the emission data need to be collected. The paper provides a discussion on what data need to be collected and how it should be processed to realize the stated goals. Established commercial platforms can be used as the gateways for data collection and processing for the logistics optimization purposes, as well as national and international

statistics bureaus for the independent data collection and processing related to the policy making process.

7. References

Bhat, C. R. (2000). FLEXIBLE MODEL STRUCTURES FOR DISCRETE CHOICE ANALYSIS. IN: HANDBOOK OF TRANSPORT MODELLING.

Blumenfeld D.E., Burns L.D., Diltz D.J. (1985), Analyzing trade-offs between transportation, inventory and production costs on freight networks, *Transportation Research Part B: Methodological*, Volume 19, Issue 5, October 1985, Pages 361–380

Campbell, J. F. (1994). Integer programming formulations of discrete hub location problems. *European Journal of Operational Research*, 72(2), 387-405.

Davydenko, I., Ehrler, V., de Ree, D., Lewis, A., & Tavasszy, L. (2014). Towards a global CO2 calculation standard for supply chains: Suggestions for methodological improvements. *Transportation Research Part D: Transport and Environment*, 32, 362-372.

Davydenko, I. Y. (2015). Logistics chains in freight transport modelling.

Davydenko, I., Nesterova, N. N., Ehrler, V., Illie, R., Lewis, A., Swahn, M., & Smith, C. (2018). Testing Results. Deliverable 4.4. TNO.

Davydenko I., M. Hopman, R.N. van Gijlswijk, A. Rondaij, J.S. Spreen (2019), Towards harmonization of Carbon Footprinting methodologies: a recipe for reporting in compliance with the GLEC Framework, Objectif CO2 and SmartWay for the accounting tool BigMile™, TNO Report 2019 R11486, 31 October 2019.

European Commission, https://ec.europa.eu/clima/policies/strategies/2030_en, consulted on 28 November 2019

Goyal, Suresh Kumar. "Economic order quantity under conditions of permissible delay in payments." *Journal of the operational research society* 36.4 (1985): 335-338.

Greene, S., Lewis, A., 2016. GLEC Framework for Logistics Emissions Methodologies. Report Global Logistics Emission Council, Smart Freight Center, Amsterdam, Netherlands.

Haug, P. (1985). A multiple-period, mixed-integer-programming model for multinational facility location. *Journal of Management*, 11(3), 83-96.

Hekkenberg, Michiel, and Robert Koelemeijer, eds (2018). Analyse van het voorstel voor hoofdlijnen van het klimaatakkoord. PBL Planbureau voor de Leefomgeving.

Klimaatakkoord (2019), Den Haag

Melo M.T., Nickel S., Saldanha-da-Gama F. (2009), Facility location and supply chain management – A review, *European Journal of Operational Research*, Volume 196, Issue 2, 16 July 2009, Pages 401–412

Montreuil, B. (2011). Toward a Physical Internet: meeting the global logistics sustainability grand challenge. *Logistics Research*, 3(2-3), 71-87.

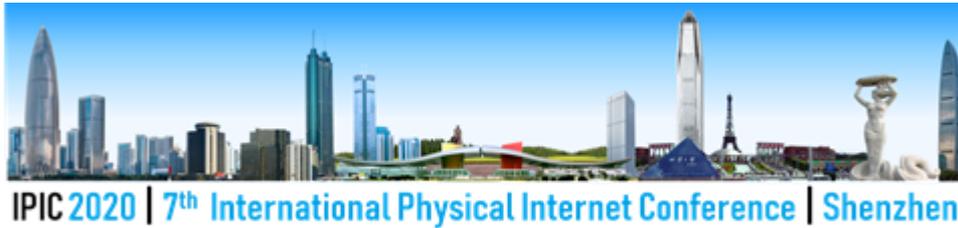
Moses, L. N. (1958). Location and the theory of production. *The Quarterly Journal of Economics*, 72(2), 259-272.

Smokers, R.T.M., Davydenko, I., Kok, R. and Spreen, J.S., Sustainable logistics (2019), Roadmapping and carbon footprinting as tools for realization of environmental goals. TNO 2019 R10104, April 2019.

Tavasszy, L., Minderhoud, M., Perrin, J. F., & Notteboom, T. (2011). A strategic network choice model for global container flows: specification, estimation and application. *Journal of Transport Geography*, 19(6), 1163-1172.

Tavasszy, L., Behdani, B., & Konings, R. (2017). Intermodality and synchromodality. In *Ports and Networks* (pp. 251-266). Routledge.

van Riessen, B., Negenborn, R. R., & Dekker, R. (2015, September). Synchromodal container transportation: an overview of current topics and research opportunities. In *International conference on computational logistics* (pp. 386-397). Springer, Cham.



Design and Evaluation of Routing Artifacts as a Part of the Physical Internet Framework

Steffen Kaup^{1,4}, André Ludwig², Bogdan Franczyk^{1,3}

1. Leipzig University, Information Systems Institute, Leipzig, Germany
2. Kühne Logistics University, Computer Science in Logistics, Hamburg, Germany
3. Wrocław University of Economics, Wrocław, Poland
4. Mercedes-Benz AG, Group Research and Sustainability, Böblingen, Germany

*steffen.kaup@uni-leipzig.de; bogdan.franczyk@wifa.uni-leipzig.de;
andre.ludwig@the-klu.org*

Keywords: *Routing Artifacts for Physical Internet, Vehicle Mesh Networks,
Multi Agent Logistics Cloud, Physical Internet Framework*

Abstract

“Global freight demand will triple between 2015 and 2050, based on the current demand pathway”, as predicted in the Transport Outlook 2019 (Forum and International Transport Forum, 2019, p. 36). Based on the current traffic situation in the existing transport infrastructure, an increase in traffic on this scale is hardly conceivable. Hence, a revolutionary change in transport efficiency is urgently needed. One approach to tackle this change is to transfer the successful model of the Digital Internet for data exchange to the physical transport of goods: The so-called Physical Internet (PI, or π). The potential of the Physical Internet lies in dynamic routing, which increases the utilization of transport modalities, like trucks and vans, and makes transport more efficient. The main physical entities in the Physical Internet include π -nodes, π -containers and π -transporters. Previous concept transfers have identified and determined the π -nodes as routing entities. Here, the problem is that the π -nodes have no information about real-time data on transport vacancies. This leads to a great challenge for the π -nodes with regard to routing, in particular in determining the next best appropriate node for onward transport of the freight package. In the near future, it can be assumed that series production vehicles or vehicle connected devices (Tran, Tran and Nguyen, 2014) will have real-time information about their load utilization. In current pre-series π -transporters the workload and thus the available space is detected either via RFID/NFC¹ technology (De Wilde, 2004), Bluetooth (Meller, Ward and Gesing, 2020), motion sensors (Knuepfer, 2007) or camera systems (Calver, Cobello and McKenney, 2008). This paper evolved the state of research concept as an artifact that considers the π -nodes as routers in a way that it distributes and replicates real-time data to the π -nodes in order to enable more effective routing decisions. This real-time data is provided by vehicles, or so-called π -transporters, on the road. Therefore, a second artifact will be designed in which

¹ RFID/NFC: Radio-Frequency Identification / Near-Field Communication

π -transporters take over the routing role. In order to be able to take a holistic perspective on the routing topic, the goods that are actually to be moved, the so-called π -containers, are also designed as routing entities in a third artifact. These three artifacts are then compared and evaluated for the consideration of real-time traffic data. This paper proposes π -transporters as routing entities whose software representatives negotiate freight handover points in a cloud-based marketplace. The implementation of such a marketplace also allows the integration of software representatives for stationary π -nodes, which contribute their location and capacity utilization levels to the marketplace. The result makes a valuable contribution to the implementation of the routing component as a part of the Physical Internet framework.

1. Introduction

The demand for transport will continue to rise strongly in the coming decades. “*Global freight demand will triple between 2015 and 2050, based on the current demand pathway*”, as predicted in the Transport Outlook 2019 (Forum and International Transport Forum, 2019, p. 36). The majority of goods in Germany are transported by trucks. This corresponds to about 72 percent of freight traffic in tonne-kilometres in 2018 (*Güterverkehr 2018*, 2019), of which 37 percent are empty runs (*Verkehr deutscher Lastkraftfahrzeuge*, 2018). Based on the current traffic situation in the existing transport infrastructure, an increase in traffic in the predicted scale is hardly conceivable. Hence, a revolutionary change in transport efficiency is urgently needed. One approach to tackle this change is to transfer the successful model of the Internet (in the following Digital Internet, DI) for data exchange to the physical transport of goods: The so-called Physical Internet (PI, or π). The idea of the Physical Internet is “*a vision of how physical objects might be moved via a set of processes, procedures, systems and mechanisms from an origin point to a desired destination in a manner analogous to how the Internet moves packets of information from a host computer to another host computer*” (Franklin, 2016).

With the PI, methods of a very established information network, the DI, are transferred to physical goods transport. This requires radical changes in current processes, which have an impact on the software and hardware of the involved network elements. As a basis for the following work, PI-specific terms need to be defined.

Definitions

The key physical elements, or also entities, in the Physical Internet include π -nodes, π -containers and π -transporters, as introduced by its inventor (Montreuil, 2012). In this paper, the subgroup π -transporter is considered as representative for π -transporter. To use the terms as clearly as possible, these entities are defined below.

Definition 1-1: π -node

A π -node represents a connection point in a network. A π -node is characterized by at least two connections to other network elements, such as other π -nodes. In general, a π -node has the ability to detect, process and forward transmissions for other network nodes.

Definition 1-2: π -container

A π -container encloses freight in such a way that it is made transportable according to its requirements. A π -container also has information about the type of goods, and their source and destination of transport.

Definition 1-3: π -transporter

A π -transporter is a moving vessel which enables one or more π -containers to be transported. In most cases these are road-bound vehicles such as cars, vans or trucks. Together with π -conveyors and π -handlers, π -transporters belong to the group of π -movers.

Where the Problem lies: The challenge of π -nodes as routers

Why is it relevant to re-think the routing nucleus of the PI? In the Digital Internet, billions of network nodes around the world are interconnected. Messages, consisting of data packets, are not transported along a predetermined route, but only to the nearest node, which then decides to which node it will forward the data packet next (Kaup and Neumayer, 2003). Hence, the nodes in the Digital Internet take over the routing function. Each network node has a forwarding table, which gives it the competence to identify the next best node for forwarding as shown in figure 1. These tables contain the next possible forwarding hubs and the number of steps to the destination, so-called hops.

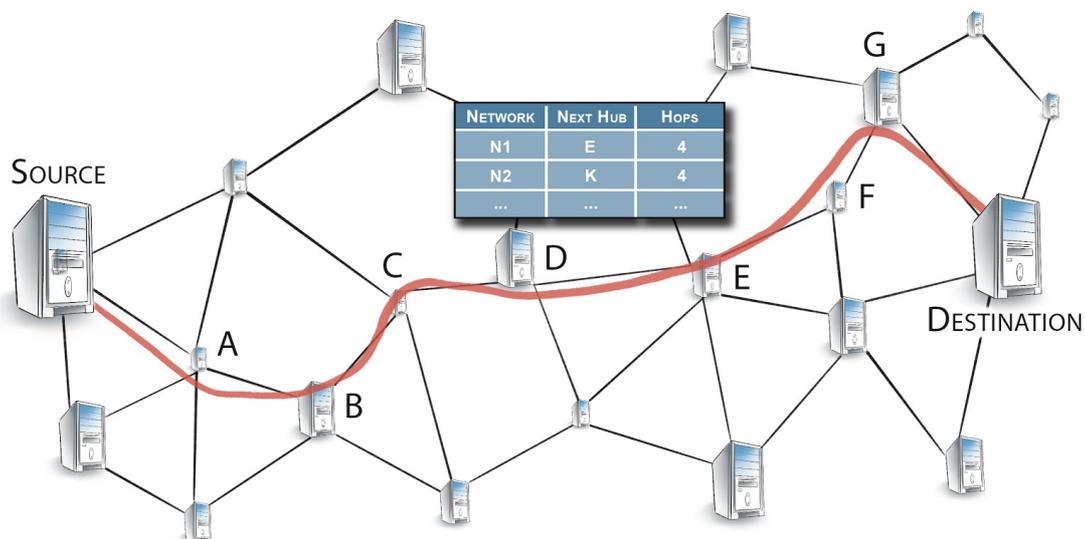


Figure 1: Routing within the Digital Internet via forwarding tables (own visualization)

The one-to-one transfer leads to the fact that nodes in the Digital Internet will correspond to π -nodes in the Physical Internet. These π -nodes, or so-called transshipment hubs, have two tasks in this concept: the physical handling of goods and the intelligence to identify the next suitable hub for further transport of the goods. A major weakness of this concept regarding the routing intelligence of the π -nodes is that they do not have any real-time traffic data. According to the current status of the concept, information such as the distance to the next transshipment hub and its accessibility will be used to identify the next step of the transport chain. These criteria are only of limited suitability for successful routing. To improve the routing competence of the hubs, real time traffic bandwidth must be considered. It can be assumed that vehicles or vehicle connected devices will have real-time information about their load utilization in the near future (Tran, Tran and Nguyen, 2014). This leads to the central Research Question (RQ) of this paper:

Where and how can the routing decision for road-based vehicles be made if real-time data about transport vacancies might be taken into account over the whole PI transport system? that is solved with the following Sub-Questions (SQ):

- SQ₁: How can the concept of π -nodes as the routing elements be extended in order to use real-time data about transport vacancies?
- SQ₂: Are there alternatives to π -nodes as routers in order to fulfill the requirement of real-time traffic data usage for routing?
- SQ₃: How valuable are these concepts in respect to their routing ability?

The structure of this paper follows the proven process of a design science approach (Wieringa, 2014). Related work and previous research are presented in Section 2. Then, Design Science Research is conducted, consisting of two elementary process steps: ‘Build’ and ‘Evaluate’. Within the Build process, activities are performed that produce design artifacts that are able to use live traffic data as input for the routing within the PI, as described in Section 3. The subsequent ‘Evaluate’ process in Section 4 evaluates the design artifacts through pre-defined success criteria and provides feedback on it (Österle *et al.*, 2011). Then, Section 5 gives a summary of the results in the form of a concluding statement and provides an outlook on recommended further research on this topic.

2. Related Research

Research related to the Physical Internet dates back to the year 2009, starting with the idea of Montreuil to transform the Digital Internet to a Physical Internet (Montreuil, 2012). Together with Ballot and Meller he wrote a textbook that describes a lot of facets of this transformation (Ballot, Montreuil and Meller, 2014). During the past years and mainly communicated through the proceedings of the International Physical Internet Conference (IPIC) many papers have been published on the idea. They also contain routing mechanisms within the PI. Furthermore, working papers from a European research group, ALICE², became available. Based on a systematic literature research regarding the PI, issues related to ‘routing mechanisms’ and any kind of ‘intelligent π -elements’, the overview in Table 1 was created.

Table 1: Results of Systematic Literature Review

	≤ 2013	2014	2015	2016	2017	2018	2019	2020
Textbooks	-	-	1	-	-	-		1
IPIC Conference Papers	-	1	1	1	1	2	2	
Science Direct	1	-	-	2	-	-	3	
SpringerLink	2	1	2	2	-	2	3	1
Elsevier	3	-	1	-	-	-	-	-
Emerald Insight	-	-	-	-	1	1	-	1
Cornell University	-	-	-	-	-	-		2
Working papers ALICE	-	-	-	-	1	-	-	1
Patents	-	-	-	-	1	-	-	-
Other publications	2	-	-	1	2	-	1	-
Total	8	2	5	6	6	5	9	≥ 6

² Alliance for Logistics Innovation through Collaboration in Europe

All these sources describe the routing function as an abstract layer or assume that the hubs, or so-called π -nodes, in the PI will take over this function (Ballot, Gobet and Montreuil, 2012). ‘*Along the whole transportation process through the logistics network, the loading unit is connected and interacts with the different logistics nodes*’ (Liesa et al., 2020, p. 37). To date, there is little thought about which criteria are suitable to determine the next best π -node. Previous research suggests that criteria such as ‘distance to the next transshipment point’, ‘reachability of the next transshipment point’ and the ‘probability of further transport’ can be used (Montreuil, 2012), but the point of efficient routing needs further research (Sternberg and Norrman, 2017). As an alternative, flow control logic and network management applications could be implemented by a cloud solution (Ballot, Montreuil and Meller, 2014). As recommended in (Becker, 2012), patents were also included in the literature search. One patent (Kaup, 2017b) describes a holistic cloud solution that holds information about all traffic bandwidth and related vacant cargo space and thus gives the freight container and its vehicle a signal when a modality change is considered suitable. Although this takes place locally at π -nodes, it is controlled by freight container representatives within a traffic cloud. Bandwidth information might be collected from π -transporters on the road (Kaup and Demircioglu, 2017). Due to this alternative, scientific databases were searched for the term ‘cloud logistics’ and the results were included in the ‘Build’ and ‘Evaluation’ phase (Ehrenberg and Ludwig, 2014) (Glöckner, Ludwig and Franczyk, 2017) (Ludwig, 2014). This paper extends the current solution ‘ π -nodes as routers’ and adds other perspectives to the question of where to position routing intelligence. It proposes to use the π -transporters as routers in combination with a cloud-based virtual marketplace.

3. Design of Entities for Routing

Design science research is about artifacts in a context (Wieringa, 2014). In the following subsections, each of the elementary entities were put as design artifacts in the context of the routing role. Based on the insights gained in the analysis of existing research, the existing concept of ‘ π -nodes as routers’ will be extended by the integration of real traffic data into the routing decision process. Two further artifacts were as alternatives to the extension of the existing approach. The ‘Build’ process of the designed artifacts was gained through a joint workshop at ‘Mercedes-Benz Innovation Studio’ with engineers and researchers in the automotive sector, the field of communication networks and in the world of logistics and telematics.

3.1 Extension of the concept π -nodes as routers

The entity π -nodes in the role as routers corresponds most likely to a one-to-one transfer from the routing mechanism of the Digital Internet to the world of physical objects. On the Digital Internet, a distinction is made between static and dynamic routing (Badach and Hoffmann, 2019). In the original context, static routing specifies a defined route definition for data exchange across different nodes. In the physical world, this can be compared with intermodal contract logistics that follow determined ways. This type of routing therefore already seems to be well implemented in the world of physical objects. The other type of routing on the DI is the dynamic routing. Dynamic routing means that the network nodes are responsible for finding the best route for individual data packages. This is done on the basis of criteria such as cost or transmission time. In order to do this, the network nodes must have knowledge of the cost or transmission time of the respective partial routes among themselves. This routing knowledge information is replicated to the network nodes via the so-called Border Gateway

Protocol (BGP) (Badach and Hoffmann, 2019) as described in the introduction. The π -nodes of the Physical Internet correspond to the network nodes in the Digital Internet. In this model, transshipment points, such as rest stops and forwarding agents' yards, decide how the journey might continue for a π -container. The π -transporters (vehicles) would follow instructions set by the π -nodes, transmitted to them e.g. via a fleet management system (Liesa et al., 2020). For this, the π -nodes first need information about which possible routes are available or which possible π -transporters still have free capacity on these routes. This information is collected from the π -transporters.

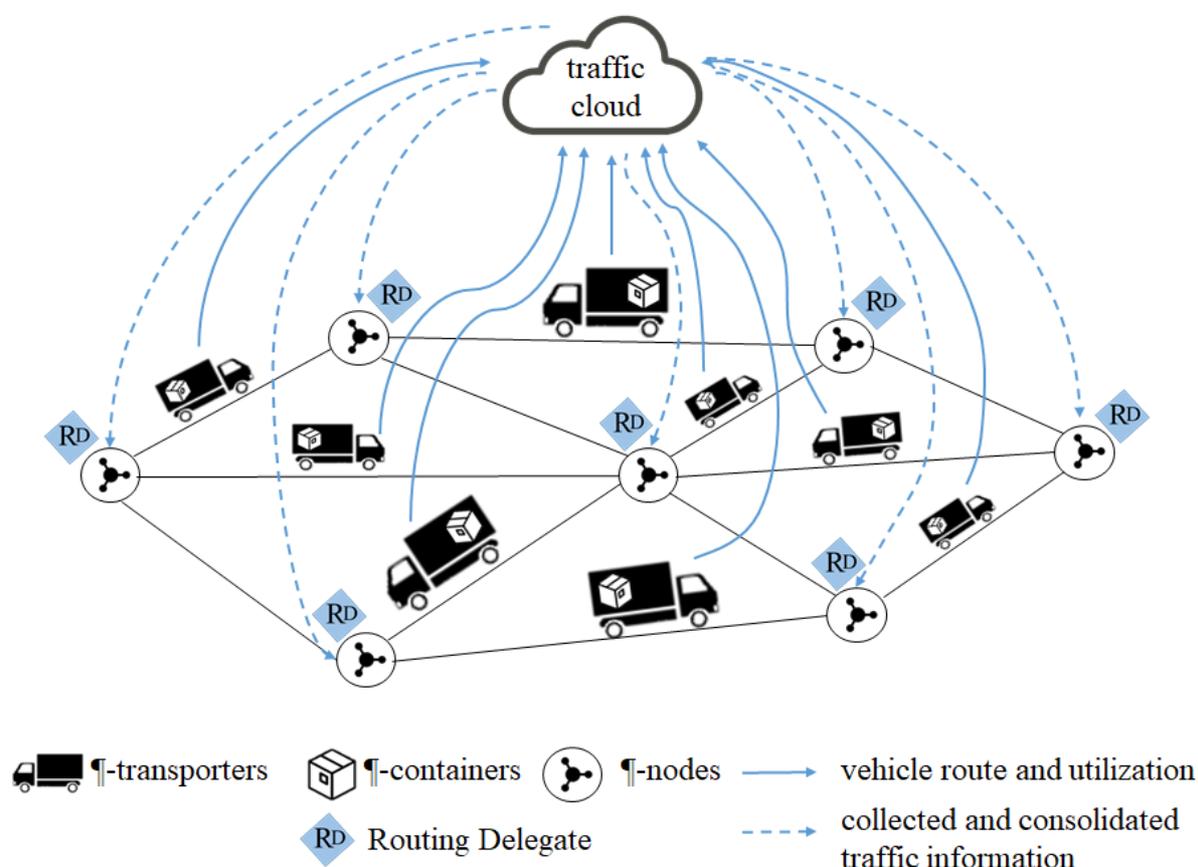


Figure 2: System of π -nodes empowered through traffic relevant data

They would either have to transmit this data to all surrounding π -nodes or send it to a common traffic cloud, which replicates the information to the π -nodes in analogy to the BGP. The routing decision is then finally made by the π -nodes. The π -nodes can be seen as Routing Delegates (R_D) of the traffic cloud (Gamma, 1995). Figure 2 visualizes an example of a cloud-supported system with ' π -node delegates' as routers.

3.2 Alternative Entity π -transporters as routers

With π -transporters as routers, the vehicles on the road are in the role of decentralized real-time decision making. They have knowledge about their planned routes and ideally about their load conditions. Tracking of the load status of a vehicle can be realized via a small onboard network, e.g. RFID/NFC (De Wilde, 2004), Bluetooth (Meller, Ward and Gesing, 2020), motion sensors (Knuepfer, 2007) or camera systems (Calver, Cobello and McKenney, 2008). This is visualized in Figure 3 by displaying a small network symbol with three arcs in

the vehicles. Also, transporters are marked as Routing-Hubs (R_H) in this Figure. Through connecting to each other via car-to-x communication, or so-called mesh networks (Jiang *et al.*, 2020), π -transporters are able to exchange relevant information with each other in order to negotiate freight exchange points among themselves. Such a vehicle network can be seen as a kind of ‘trading venue’ where next appropriate routes and necessary modality switches can be negotiated. This network might be implemented in a distributed way among the vehicles within the vehicles mesh network, e.g. using Distributed Ledger Technologies (DLT), such as blockchain (Mollah *et al.*, 2020).

This network would also be able to heal freight transport routes in case one or more of the vehicles would fail or are delayed in a traffic jam. Information on the location of hubs is not prone to change as frequently as that of traffic, so that it can be integrated into existing map material with acceptable effort. Through this technology, dynamic transfer points outside the range of regular hubs might also be negotiated among the vehicles within the mesh network.

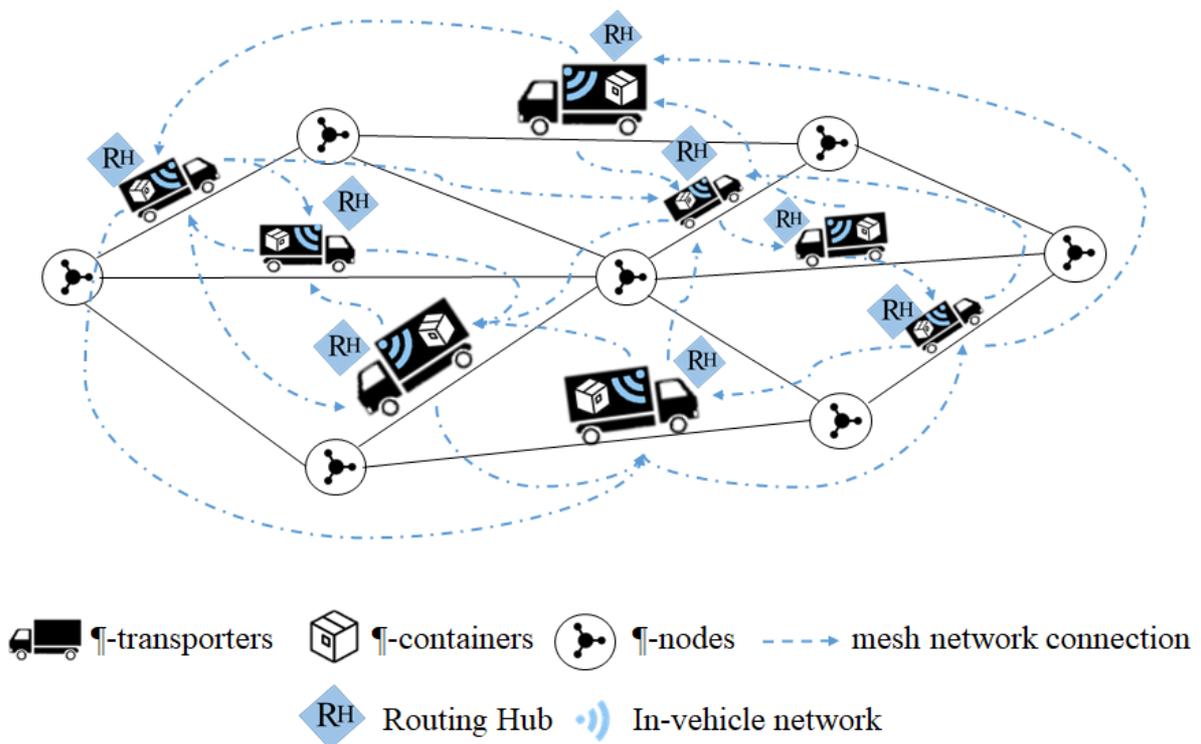


Figure 3: Example of a mesh network of π -transporters in the role of routing hubs

Hence, these transfer points do not necessarily have to correspond to static transshipment yards, but also to mobile hubs or dynamic rendezvous points. This makes the system highly flexible and efficient. In a final expansion stage, freight could already be exchanged during the ride, similar to some approaches under study for passenger transport. This routing artifact is easily scalable, so that more and more π -transporters might be enabled as mobile hubs. These vehicles trade among themselves for more or less capacity and thus earn additional money during operation. It can be said that this will enable tasks to be taken over by the π -transporters that were previously performed by freight exchange companies.

3.3 Alternative Entity π -containers as routers

To have π -containers as the routing entity would mean that the containers itself could make routing decisions. However, for this they need information about traffic conditions and

unused capacities within π -transporters. Even if they could negotiate directly with vehicles, quasi like a hitchhiker, the onward journey would not be secured. For this reason, every π -container has a software agent that represents himself in a common traffic cloud, like a digital twin (Hofmann and Branding, 2019). Hence, the traffic cloud, also called the Routing Brain (R_B), orchestrates the utilization of the π -containers in a holistic way (Kaup, 2017b), as visualized in Figure 4.

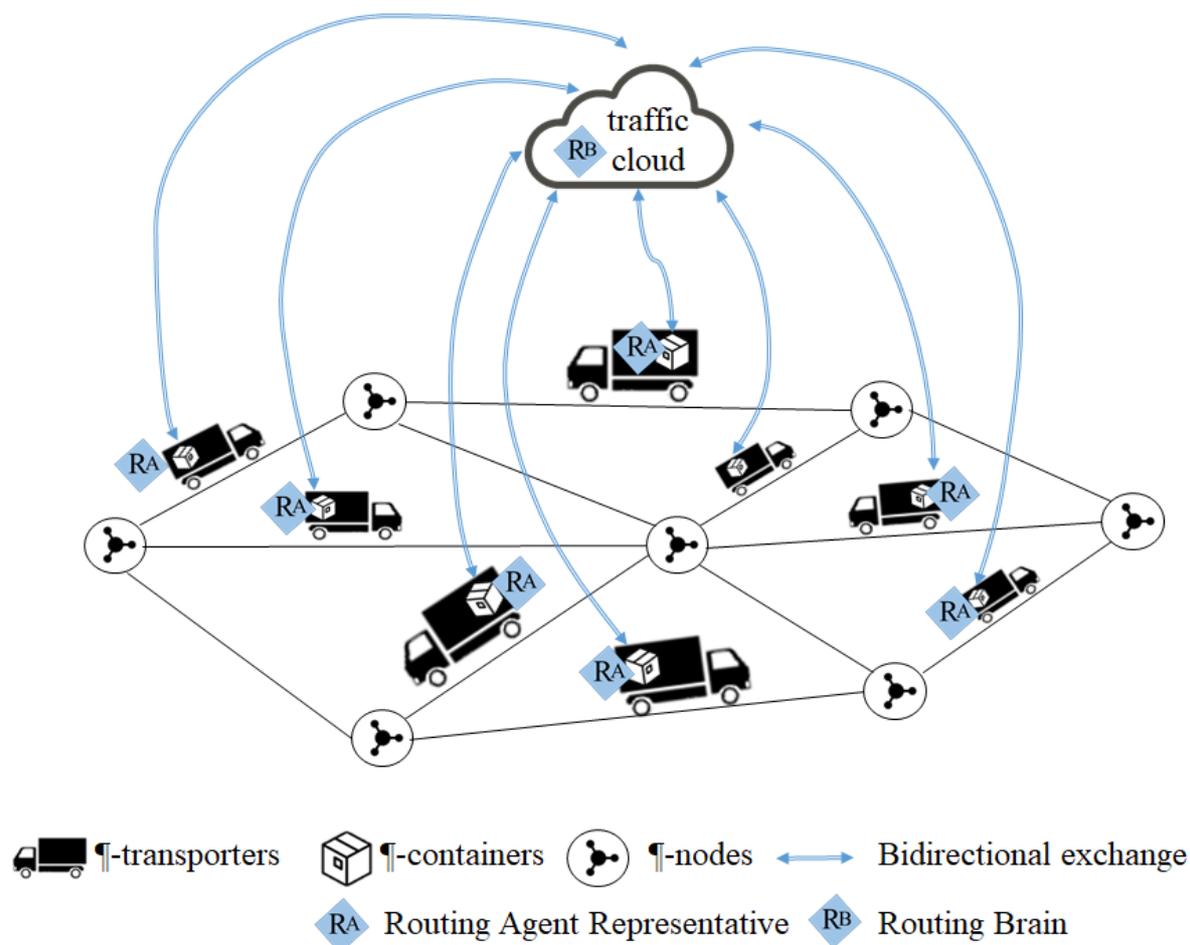


Figure 4: System of π -containers and their Routing Agents

A bidirectional exchange between π -containers and the traffic cloud is necessary in order to inform the traffic cloud about slow-moving traffic or possible accidents, which requires a network with high bandwidth and ultra-low latency (5G). Possible lacks or dysfunctionality of traffic could be identified on feedback from π -containers to the traffic cloud concerning their current location information.

This approach corresponds to the mobility behavior of people who book common modes of transport or mobility via an app (e.g. to reserve a seat on a train or a taxi on call). By using a platform, e.g. Moovel³, transport requirements are entered and then suitable intermodal routes are calculated by the cloud platform depending on time and cost. In a similar logic, mobile π -containers could be routed and operated via apps (Tran-Dang and Kim, 2018). Since containers themselves cannot enter data into an app, they need a software agent as a representative in the cloud (R_A , see also Figure 4) (Zhou and Lou, 2012). Such a container representative communicates with the π -container and steers it through the traffic network

³ Moovel Group GmbH (new name: REACH NOW), Software Company, Stuttgart, Germany

using the cloud information as the Routing Brain. For this purpose, each π -container must be equipped with a transmitting and receiving unit. Clear standards are required to integrate as many π -containers as possible in this system. The high-quality and expensive π -containers must also be reused. If possible, no empty π -container should go back anywhere. In the case of mixed operation, i.e. if existing dedicated traffic should also be used, the cloud must also have information on vehicle utilization levels, similar to π -nodes as routers.

In summary it can be said that all artifacts make routing decisions based on real-time data about transport vacancies. This kind of data is collected and made available by π -transporters in all solutions. The differences of the artifacts lie in which π -element the routing decision is made and whether this is done peer-to-peer or by software-representatives in a cloud solution.

4. Routing Entity Evaluation

In this Section, the ‘Evaluate’ process takes part, that means the designed three artifacts have been assessed in their routing context. First, evaluation criteria were determined. The criteria make it possible to assess the robustness and rigor of the designed artifacts as well as the efficiency gain and accessibility to other providers. Then, the artifacts were validated against these criteria by an empirical study.

4.1 Criteria and Methodology for Evaluation

The objective of evaluation methodology is to make sure that satisfactory progress is being made towards fulfilling deliverables and reaching relevant contributions to the PI. In literature, evaluation ‘serves the purpose of deciding whether or not to acquire or develop a technology, or the purpose of deciding which of several competing technologies should be acquired or adopted (Venable, Pries-Heje and Baskerville, 2016, pp. 77–89)’. The method in process used in this paper is the qualitative analysis of expert interviews. In order to be able to compare the artifacts with one another as well as possible, the authors chose the deductive category application (Mayring, 2014). In order to perform this and to ensure the rigor of the research, accepted criteria are needed to assess how well the constructed artifacts fit into the routing context (Österle *et al.*, 2011). But how to find accepted criteria? The reason for doing research on the PI is the respected outcome of a higher degree of ‘Efficiency’ which qualifies this criterion for the catalogue. In addition, the research group ALICE worked out ‘Seamlessness’ and ‘Scalability’ as important requirements for logistics networks (Liesa *et al.*, 2020). A seamless transition from one mode of transport to another is necessary for freight that requires special handling, e.g. refrigerated food or medicine. Scalability indicates how easy the transport network can be extended, e.g. by new routing elements. Discussions with experts from freight forwarding companies lead to the finding that scalability is not sufficient. They require interoperability of different players cross-brand. As an example, the technology company Apple makes it easy for users to add new Apple components within its own product environment, but as hard as possible to add components from competitors. To enable a great possible effect of the PI, we need the interoperability of different players and companies. For this, the criteria ‘Openness’ was added. Discussions with customers led to the finding that the newly designed system must be at least as good as current-world logistics. To ensure this, the criteria ‘Costs’, ‘Time’ and ‘Reliability’ were added, as shown in Table 2.

Table 2: Criteria for Routing Entity Evaluation

Criterion	Description
Scalability	Scalability indicates how easy the transport network can be extended, e.g. by new routing elements.
Openness	Ensures cross-brand accessibility of π -elements and the interoperability of different players, such as large shipping companies, smaller vehicle fleets, solo entrepreneurs.
Efficiency	Indicates how efficiently transport and routing take place. This is done by estimating the average capacity utilization rate of the modes of transport.
Costs	The cost of implementing a functional self-routing system consisting of the necessary adaptations to the relevant π -elements, such as π -nodes, π -transporters or the cost of developing a necessary cloud platform.
Time	Estimated qualitative transport time of representative end-to-end connections of a standardised container within the network.
Reliability	Reliability of the onward transport of goods from one hub to the next hub.
Seamlessness	The retail customer is not affected by freight changes in modes and routes due to the comprehensive and fully interconnected network. In addition, cold chains can also be ensured by reducing the probability of goods being detained during cargo handling.

In order to achieve the greatest possible gain in knowledge, a qualitative empirical approach with experts was chosen. For this, the designs of the three artifacts were discussed with experts, who imagine how such an artifact will interact with the problem context of routing freight within a Physical Internet. Then, they predicted what effects regarding the determined criteria they think this would have.

4.2 Results of Evaluation

Evaluation by expert opinion only works if the experts understand the artifacts, imagine realistic problem contexts, and make reliable predictions about the artifacts in context. Hence, it was not a trivial task to find the right experts to evaluate the designed artifacts. The requirements for the interview partners were that they had to have both expert knowledge in functioning of the DI as well as in transport logistics. The central statements of these experts (N=9) are shown in a distilled form in Table 3. The interviewed experts were divided into the categories ‘professors or chairs of renowned universities’^U, ‘founder of highly innovative startups or CEO’s of consulting companies’^C and ‘research leaders within the automotive industry’^A. The superscripts (U, C, A) on the central statements in the evaluation matrix indicate the category to which the expert who made this statement belongs.

Table 3: Evaluation Matrix of Routing Entity Artifacts

Criterion	Routing Entity #1: π -nodes	Routing Entity #2: π -transporters	Routing Entity #3: π -containers
Scalability	<ul style="list-style-type: none"> + As soon as a traffic cloud is implemented, more and more π-nodes can be added^U - π-nodes still must be provided with relevant traffic information^U 	<ul style="list-style-type: none"> + Most vehicles already have communication technology on board^A - Limited suitability for the long haul because of limited communication range to other vehicles^C 	<ul style="list-style-type: none"> + Once an infrastructure is agreed, π-containers can be added easily^U - A global communication standard is needed for all containers^C - 5G is required^U
Openness	<ul style="list-style-type: none"> + Other modalities are convenient to include, because stationary hubs are often located at railway stations or ports^C - Most of π-nodes are privately owned, agreements of use are difficult to conclude^A 	<ul style="list-style-type: none"> + New routing elements in form of π-transporters can be easily integrated^A - Building Ad-hoc networks with other modalities, like trains or ships is seen as a challenge^C 	<ul style="list-style-type: none"> + Good, under the condition that the cloud-platform is open and barrier-free for all kinds of π-container representatives^U - In a mixed operation with existing traffic, an additional connection from π-transporters to the cloud is required^A
Efficiency	<ul style="list-style-type: none"> - Only static π-nodes can be used^U 	<ul style="list-style-type: none"> + Dynamic π-nodes possible (rendezvous points)^A 	<ul style="list-style-type: none"> + Dynamic π-nodes possible (rendezvous points)^A
Costs	<ul style="list-style-type: none"> - Collecting traffic information and replicating them to π-nodes is complex^U 	<ul style="list-style-type: none"> + Many π-transporters already have telecommunication systems, it is “just” a software topic^A 	<ul style="list-style-type: none"> - Expensive, as each π-container would have to be equipped with a long-range telematic unit^C
Time	<ul style="list-style-type: none"> o Can only be evaluated after implementation or simulation^{U,C,A} 		
Reliability	<ul style="list-style-type: none"> + The responsibility clearly lies with the π-nodes. Hence, reliability is seen as high.^C 	<ul style="list-style-type: none"> o Reliability depending on defined communication and negotiation standards^A 	<ul style="list-style-type: none"> o Reliability depending on Software Agents in the Cloud and the integration of existing traffic^A
Seamless-ness	<ul style="list-style-type: none"> - The π-nodes must get information about the transport network from somewhere^U 	<ul style="list-style-type: none"> + Good, if manufacturer-independent standard exists and DLT platform works^A 	<ul style="list-style-type: none"> + Very good, if non-proprietary container-standard exists^C

The findings of the designed artifacts indicate that each of the solutions has advantages and disadvantages. The most technically mature solution is not always the one that can be quickly

and safely established on the market. In the following subsections the obstacles and opportunities of the routing artifacts are discussed.

Opportunities and obstacles of ‘ π -nodes as routers’

All the interviewed experts agree that it is necessary to include current information on traffic volume and vehicle utilization in the routing process. The artifact ‘ π -nodes as routers’, as extended in Section 3.1, rises and falls with the ability to replicate traffic and vehicle related information on the π -nodes. Also, this entity design is limited to stationary hubs as routing elements. This may seem simple at a first glance, but it should be noted that most of the π -nodes are privately owned and operated, because almost all logistic hubs are dedicated to a freight forwarding company, e.g. DHL. It must be ensured that these possible π -nodes can also be used without manufacturer discrimination, perhaps even with political support. With π -transporters or π -containers as routers, dynamic transfer points are theoretically possible. But, how does the goods turnover look like there? Since there are no stationary handling robots on site of π -nodes, the physical handling of goods either has to be carried out by humans or the π -containers have to be mobilized in some way, for example through mobile delivery robots (Canoso, Binney and Rockey, 2017).

Opportunities and obstacles of ‘ π -transporters as routers’

Six of the nine experts surveyed, particularly those from the automotive industry, are convinced that the concept of smart vehicles, as introduced in Section 3.2, would ease the issue of protocol development and complexity quite considerably. In the artifact ‘ π -transporters as routers’, π -transporters (vehicles) negotiate the next best mode of transport for containing freight among themselves, just as autonomous vehicles will have to negotiate the right of way with each other in the future. It has the charm that (almost) every vehicle (as the most common implementation of π -transporters) has a connection to the Internet and therefore to another vehicle. If the vehicle itself should not have this kind of connection, there is often a driver sitting in it, who has a connection to the Internet via personal devices. Hence, π -transporters could be connected to each other easily, e.g. via an app as described in (Tran, Tran and Nguyen, 2014). In order to negotiate the onward transport of goods between vehicles, they must have information about the π -containers and their destinations. If the π -transporters are intelligent, then an entirely different concept for control can be utilized that really simplifies the replication problem, resulting from the concept of π -nodes as routers. They still need to understand the state of the network, but rerouting and load management could be organized by the vehicles themselves, e.g. by using a blockchain-backed broker platform (Meyer, Kuhn and Hartmann, 2019). Tracking and tracing of π -containers within the vehicles can be realized through technologies like RFID (De Wilde, 2004) or motion sensors (Knuepfer, 2007). Hence, the π -transporters act as a ‘kind of mobile hub’ with acceptable efforts and costs. By using exchangeable body capsules, the π -transporters entity concept would also be transferable to passenger transportation and mobility solutions (Froböse, 2012).

Opportunities and obstacles of ‘ π -containers as routers’

The concept of ‘ π -containers as routers’ would be as if packets on the Internet were made intelligent by their cloud representatives and could, therefore, dynamically manage their movements through the network. In the eyes of many experts, this concept requires a complex sending and receiving unit at the side of the π -containers, which must be able to connect to their cloud representatives. Likewise, the network development for 5G must be sufficiently progressed. This requires much effort and leads to high costs. In a mixed

operation, the cloud must first be supplied with holistic traffic data, as described in Section 3.1. The routing algorithm must also ensure that the system does not return these expensive π -containers empty again. As an alternative approach, π -containers could also become an integral part of mesh-networks, like mentioned in (Ballot, Montreuil and Meller, 2014) as follows: ‘*A container that becomes an integral part of the Internet of Things, along with its handling and storage equipment and transportation, then allows these to be coordinated through machine-to-machine communication.*’ But, it will take a long time to develop a smart container network architecture across different manufacturers of π -containers (Marino *et al.*, 2019) or to connect π -containers directly to π -transporters or π -hubs with ultra-low latency via 5G (Pagano *et al.*, 2019).

According to the highest number of advantages and the lowest number of disadvantages, the artifact ‘ π -transporters as routers’ is considered the most appropriate solution. Hence, this paper proposes to use π -transporters, in particular vehicles on the road, as routers. Either the vehicles communicate with each other via ad-hoc mesh networks or software representatives perform this task for them in a kind of logistics cloud (Glöckner, Ludwig and Franczyk, 2017). This depends on how much computing power is available in the vehicles and whether they can communicate with each other across manufacturers. Artifact ‘ π -transporters as routers’ and artifact ‘ π -containers as routers’ could be combined to overcome the weakness of the low range of the vehicle mesh network. Five of the experts consider it useful to have all traffic relevant data in one place. Hence, a cloud solution is preferred, which serves as a virtual marketplace for freight exchange between π -transporters. Most of the transshipment operations will continue to take place at stationary hubs. The implementation of a virtual marketplace also allows the integration of software representatives for stationary π -nodes, which contribute their location and capacity utilization levels to the marketplace. The opening of the marketplace for π -container representatives can also be considered, so that customers can directly contribute freight with transport needs. To better predict routes and thus improve long-distance routing, Artificial Intelligence in combination with Game Theory methods could further increase the effectiveness of the virtual marketplace. This could further optimize the multi-agent system in order to support software agents finding their best negotiation partners for taking over the freight for onward transportation.

5. Conclusions and further work

In the logistics of physical objects, a big challenge lies in non-use of free transport capacities. A method, similar to dynamic routing on the Digital Internet, can be expected to increase efficiency in the whole transport chain. Current research identified and determined the hubs in charge of routing freight efficiently through a Physical Internet. This paper addresses the as yet unsolved problem of how to identify the next best appropriate hub for onward transport based on real-time traffic data. Hence, this paper answers the central Research Question (RQ) of where and how the routing decision for road-based vehicles can be made if real-time data about transport vacancies might be taken into account for the routing schemes within the PI. To solve this problem, the existing concept of routing π -nodes was extended with the supply of real-time traffic data by collecting it from π -transporters and storing them into a cloud. This cloud provides and replicates the data about transport vacancies to the π -nodes and therefore answers SQ₁. This led to the design of an alternative artifact that sets the π -transporters directly in the routing role. In order to negotiate transfer points with other transporters, a mesh network between them was designed. Last but not least, the containers themselves were put into the routing role via software-representatives that negotiates further transport possibilities in something like a cloud-based marketplace. These two designed alternatives, π -transporters and π -containers as routers, answer SQ₂. Then, the three artifacts

were evaluated due to their routing competence regarding criteria, derived from previous literature and research, that address SQ₃. As a result, there is no solution that is the most suitable in all aspects, but the evaluation proposes π -transporters as routers. The reasons for this are twofold: the π -transporters collect the traffic data due to identify available transport bandwidth and already have suitable telematics devices on board in order to negotiate possible transfer points for freight. A key finding of this paper is that real-time data about available vacancies in π -transporters, collected by a swarm of π -transporters, is key as a reasonable basis for routing freight in an effective way through the PI. Either the vehicles communicate with each other via ad-hoc mesh networks or software representatives perform this task for them in a common logistics cloud. This depends on how much computing power is available in the π -transporters and whether they can communicate with each other across manufacturers. From a programmer's perspective, it is of decisive advantage to have all data in one place. Therefore, a cloud solution is preferred, which serves as a virtual marketplace for freight exchange between π -transporters. In the same way, this marketplace could be extended by software-representatives of π -nodes and π -containers and thus let them participate in the negotiation process of the intermodal transport chain.

References

- Badach, A. and Hoffmann, E. (2019) "Technik der IP-Netze," *Technik der IP-Netze*, pp. I–XXVIII. doi: 10.3139/9783446455115.fm.
- Ballot, E., Gobet, O. and Montreuil, B. (2012) "Physical Internet Enabled Open Hub Network Design for Distributed Networked Operations," in Borangiu, T., Thomas, A., and Trentesaux, D. (eds.) *Service Orientation in Holonic and Multi-Agent Manufacturing Control*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 279–292. doi: 10.1007/978-3-642-27449-7_21.
- Ballot, E., Montreuil, B. and Meller, R. D. (2014) *The Physical Internet: The Network of Logistics Networks*. Available at: https://books.google.com/books/about/ThePhysical_Internet.html?hl=&id=iX2ZAQAACAAJ.
- Becker, M. (2012) "Hinweise zur Anfertigung eines Literatur-Reviews," *Leipzig: Universität Leipzig*. Available at: <http://www.caterdev.de/wp-content/uploads/2013/04/reviews.pdf>.
- Calver, A. J., Cobello, R. and McKenney, K. G. (2008) "Cargo sensing system," *US Patent*. Available at: <https://patentimages.storage.googleapis.com/e7/47/f2/bc1baa0eb1262b/US7421112.pdf> (Accessed: July 14, 2020).
- Canoso, A., Binney, J. and Rockey, C. (2017) "Mobile delivery robot with interior cargo space," *US Patent*. Available at: <https://patentimages.storage.googleapis.com/e9/1a/16/38a01b00ea67e3/US9535421.pdf> (Accessed: July 10, 2020).
- De Wilde, E. (2004) "Truck cargo management RFID tags and interrogators," *US Patent*. Available at: <https://patentimages.storage.googleapis.com/56/19/d3/2841e1e584b40b/US20040069850A1.pdf> (Accessed: July 10, 2020).
- Ehrenberg, D. and Ludwig, A. (2014) "Cloud Logistics - Innovationspotenziale durch Virtualisierung von Logistikeinheiten nutzen," *Wirtschaftsinformatik & Management*, 6(1), pp. 6–9. doi: 10.1365/s35764-014-0379-7.
- Forum, I. T. and International Transport Forum (2019) "ITF Transport Outlook 2019 (Summary in

English),” *ITF Transport Outlook*. doi: 10.1787/c013afc7-en.

Franklin, R. (2016) “The Physical Internet - Just what is this Idea.” Available at: <https://transmetrics.eu/2016/content/uploads/Rod-Franklin-Physical-Internet-presentation.pdf> (Accessed: November 27, 2019).

Froböse, R. (2012) *Mein Auto repariert sich selbst: Und andere Technologien von übermorgen*. John Wiley & Sons. Available at: <https://play.google.com/store/books/details?id=9oqjYka5TmYC>.

Gamma, E. (1995) *Design patterns: elements of reusable object-oriented software*. Pearson Education India. Available at: http://asi.insa-rouen.fr/enseignement/siteUV/genie_logiciel/supports/ressources/exemples_de_la_vie_reelle_pour_illustrer_pattern.pdf.

Glöckner, M., Ludwig, A. and Franczyk, B. (2017) “Go with the flow-design of cloud logistics service blueprints,” in *Proceedings of the 50th Hawaii International Conference on System Sciences*. Available at: <https://scholarspace.manoa.hawaii.edu/handle/10125/41776>.

Güterverkehr 2018 (2019). Statistisches Bundesamt. Available at: https://www.destatis.de/DE/Themen/Branchen-Unternehmen/Transport-Verkehr/Gueterverkehr/_inhalt.html#sprg235064.

Hofmann, W. and Branding, F. (2019) “Implementation of an IoT- and Cloud-based Digital Twin for Real-Time Decision Support in Port Operations,” *IFAC-PapersOnLine*, pp. 2104–2109. doi: 10.1016/j.ifacol.2019.11.516.

Jiang, X. *et al.* (2020) “Hybrid Low-Power Wide-Area Mesh Network for IoT Applications,” *arXiv [cs.NI]*. Available at: <http://arxiv.org/abs/2006.12570>.

Kaup, S. (2017a) “Impact of the Physical Internet on Sustainable Logistics and Transportation.” *Future of Transportation World Conference*, 5 July.

Kaup, S. (2017b) “Verfahren zur Planung und Ermittlung einer optimalen Transportroute,” *Patent*.

Kaup, S. and Demircioglu, A. V. (2017) “Von der Crowd-Logistik hin zu einem ganzheitlichen Ansatz hocheffizienten Warentransports,” *Wirtschaftsinformatik & Management*, 9(3), pp. 18–27. doi: 10.1007/s35764-017-0052-z.

Kaup, S. and Neumayer, B. (2003) *Rechnernetze und Datensicherheit*. Shaker.

Knuepfer, J. (2007) “System and Method for Monitoring the Cargo Space of a Transportation Device,” *US Patent*. Available at: <https://patentimages.storage.googleapis.com/a0/aa/c2/cb2c48595f0db5/US20070241897A1.pdf> (Accessed: July 10, 2020).

Liesa, F. *et al.* (2020) “Physical Internet Roadmap.” SENSE Project (D2.1).

Ludwig, A. (2014) “Engineering und Management kundenindividueller Logistikdienste nach dem Cloud-Prinzip,” *Wirtschaftsinformatik & Management*, 6(1), pp. 46–55. doi: 10.1365/s35764-014-0384-x.

Marino, F. *et al.* (2019) “IoT enabling PI: towards hyperconnected and interoperable smart containers,” in *Proceedings of 6th International Physical Internet Conference 2019*, pp. 349–362. Available at: <https://www.iconetproject.eu/wp-content/uploads/2019/08/IPIC2019-Final.pdf>.

Mayring, P. (2014) “Qualitative content analysis: theoretical foundation, basic procedures and software solution.” AUT. Available at: https://www.ssoar.info/ssoar/bitstream/handle/document/39517/ssoar-2014-mayring-Qualitative_content_analysis_theoretical_foundation.pdf.

Meller, P., Ward, J. and Gesing, B. (2020) *Next-Generation Wireless in Logistics*. DHL Trend Research. Available at: <https://www.dhl.com/global-en/home/insights-and-innovation/thought-leadership/trend-reports/next-generation-wireless.html#:~:text=The%20latest%20DHL%20Trend%20Report,in%20the%20world%20of%20logistics>.

Meyer, T., Kuhn, M. and Hartmann, E. (2019) “Blockchain technology enabling the Physical Internet: A synergetic application framework,” *Computers & Industrial Engineering*, 136, pp. 5–17. doi: 10.1016/j.cie.2019.07.006.

Mollah, M. B. *et al.* (2020) “Blockchain for the Internet of Vehicles towards Intelligent Transportation Systems: A Survey,” *arXiv [cs.CR]*. Available at: <http://arxiv.org/abs/2007.06022>.

Montreuil, B. (2012) “Physical Internet Manifesto,” Version 1.11.1, pp. 2010–2004.

Österle, H. *et al.* (2011) “Memorandum on design-oriented information systems research,” *European Journal of Information Systems*. Taylor & Francis, 20(1), pp. 7–10. doi: 10.1057/ejis.2010.55.

Pagano, P. (2019) “The 5G-based Model-Driven real Time Module for General Cargo Management,” in. *International Physical Internet Conference 2019*.

Sternberg, H. and Norrman, A. (2017) “The Physical Internet – review, analysis and future research agenda,” *International Journal of Physical Distribution & Logistics Management*, pp. 736–762. doi: 10.1108/ijpdlm-12-2016-0353.

Tran-Dang, H. and Kim, D. (2018) “An Information Framework for Internet of Things Services in Physical Internet,” *IEEE Access*, 6, pp. 43967–43977. doi: 10.1109/ACCESS.2018.2864310.

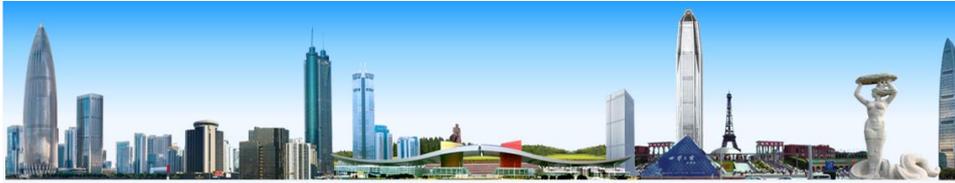
Tran, P. V., Tran, T. V. and Nguyen, H. T. (2014) “Visualization-Based Tracking System Using Mobile Device,” in Sobecki, J., Boonjing, V., and Chittayasothorn, S. (eds.) *Advanced Approaches to Intelligent Information and Database Systems*. Cham: Springer International Publishing, pp. 345–354. doi: 10.1007/978-3-319-05503-9_34.

Venable, J., Pries-Heje, J. and Baskerville, R. (2016) “FEDS: a Framework for Evaluation in Design Science Research,” *European Journal of Information Systems*. Taylor & Francis, 25(1), pp. 77–89. doi: 10.1057/ejis.2014.36.

Verkehr deutscher Lastkraftfahrzeuge (2018). Kraftfahrt-Bundesamt. Available at: https://www.destatis.de/DE/Themen/Branchen-Unternehmen/Transport-Verkehr/Gueterverkehr/_inhalt.html#sprg235064 (Accessed: November 2, 2020).

Wieringa, R. J. (2014) *Design Science Methodology for Information Systems and Software Engineering*. Springer. Available at: <https://play.google.com/store/books/details?id=xLKLQBQAAQBAJ>.

Zhou, L. and Lou, C. X. (2012) “Intelligent Cargo Tracking System Based on the Internet of Things,” in *2012 15th International Conference on Network-Based Information Systems*, pp. 489–493. doi: 10.1109/NBiS.2012.127.



Place your logo here

IPIC 2020 | 7th International Physical Internet Conference | Shenzhen

Complexity of rules in crowdsourced deliveries and its level of intrusiveness on participants: An experimental case study in the Netherlands

Xiao Lin¹, Yoshinari Nishiki², Lóránt A. Tavasszy¹

Department of Transport & Planning, Delft University of Technology, The Netherlands
Technology of Future Utopia (TOFU)

Corresponding author: x.lin(a)tudelft.nl

Abstract: *Crowdsourced logistics systems are receiving increasing attention in both industry and academia. This paper takes a unique standpoint in studying the relations between overall system performance and intrusiveness to each individual's daily lives, by performing a case study in The Hague. Volunteering cyclists were asked to transport small parcels while simulating their daily commuting behavior. Movements of parcels were recorded by GPS trackers and later analyzed. The results show that a crowdsourced logistics system can balance the overall system performance and level of intrusiveness on each participant by having well designed organizing mechanisms.*

Keywords: *Physical Internet, self-organizing system, crowdsourced logistics, participant behavior, intrusiveness*

1 Introduction

With the developing technology and society's growing concern over environment, academic and industrial communities are rethinking the way we organize production, transportation, and logistics. Among the emerging conceptual solutions, crowdsourced delivery is receiving increasing attention. Crowdsourced delivery can be defined as a delivery service, a business mode that designates the outsourcing of logistics to a crowd, while achieving economic benefits for all parties involved (Devari et al. 2017).

The concept of crowdsourced logistics is related to the trend of sharing economy and Physical Internet (Rai et al. (2017), the idea that physical objects are transported in modular packets as efficient as possible to their destinations, regardless of the route followed). By making use of crowdsourced transportation capacities, deliveries of goods are done without having to deploy dedicated logistics services. This means a reduced delivery cost for the owners of products and decreased impact on the environment. In this context, items can be transported in a "non-dedicated" context. These potential transportation capacities can be commuters, on which parcels "hitchhike": imagine a situation when your neighbor brings your Amazon parcel to you, because he has happened to be at a location where your parcel has been placed.

Literature reports some pilot projects of crowdsourced deliveries. Rougès et al. (2014) analyze 26 businesses run by companies and start-ups that provide platforms for crowdsourced delivery. They point out that, with the framework of Physical Internet, the potential power of crowdsourced delivery could be one of the alternative transportation solutions. Furthermore, a multi-segment multi-carrier delivery mode called "TwedEx" is discussed in Hodson (2013): people carry packages secondary to their daily lives e.g., commuting. Each package is handled from person to person based on overlaps in time and space until the package is delivered. This business model is further researched in simulated numerical studies in Sadilek et al. (2013). The analysis shows great potential in this business model, as it has remarkable speed and

coverage. The authors call for “constructing and fielding (such) services”, which could provide new insights for crowdsourced activities and business models.

The performances of crowdsourced delivery systems have been analyzed in abundance at a system level. Chen et al. (2015) develop a method for recommending tasks to mobile crowdworkers with the aim of maximizing the expected total rewards collected by all agents. Soto Setzke et al. (2017) develop a matching algorithm that assigns items to drivers for delivery, with the objective of minimizing the additional travel time apart from planned routes. Chen et al. (2016) use Taxi data in a city as reference to develop a strategy to minimize package delivery time by assigning paths to each package request. Arslan et al. (2018) study a dynamic pickup and delivery problem in order to match the delivery requirements to existing traffic flow. These articles investigate crowdsourced delivery at a system level, mostly to optimize the overall performance of the system by improving matching or task assignments.

The aforementioned articles give no attention to individual participants in such a system. As mentioned in Hodson (2013), participants from the crowd are not dedicated employees of delivery companies, thus the delivery tasks are only a side-objective apart from their daily lives. Carrying a parcel and giving it to someone will for sure introduce some degree of disruption to their normal living patterns. Naturally, the more disruption the system imposes to each of the participants, the more likely it will affect the willingness of participant in a negative way. Understanding the degree of disruption will no doubt help design crowdsourced delivery platforms.

Limited research has explicitly discussed about the degree of disruption as well as the willingness of participation. Kim et al. (2018) introduce a “Hit-or-Wait” approach in order to balance the timing when participants are matched with tasks with minimal disruptions of their existing route. Chi et al. (2018) explore the motivations of participants in contributing to crowdsourced projects. They use the app Crowdsourc from Google, which aims to acquire training data for machine learning projects. The result of their survey indicates that participants are eager to be recognized by an organization or a community, especially if the recognition can be globally known. Zheng and Chen (2017) investigate a crowdsourced task-assigning problem considering the possibility that participants may reject a task. They measure willingness of participation using probability of rejection. However, the practical meaning of this probability and how it can be derived from each participant remains unclear. Miller et al. (2017) study commuters' behavior by sending out surveys to understand how willing are people to participate as workers in crowdsourced logistics. Most of the relative research uses surveys or simulations, which are at a theoretical level.

In this paper, we use experimental case studies to investigate how crowdsourced transportation capacity can be best organized. We give attention to the relationship between the whole system and each individual. In particular, the link between effectiveness of system's overall performance and degree of disruption being brought to the participants. We define the likelihood that participating the crowdsourced activities disrupts a participant's daily lives as the *level of intrusiveness*. In order to observe this linkage, we conducted a case study in a small area in the Dutch city The Hague. Volunteers were invited to cycle in this area. Meanwhile, they were asked to form ad-hoc relays to deliver GPS-tracked mango parcels. In Section 2, we briefly look at the elements of self-organized, crowdsourced systems, and explain experiment design for the case study. In Section 3, the results of the experiments are analyzed. We also discuss the experiment results and the potential of this form of crowdsourced logistics. Section 4 concludes this article and points out future research directions.

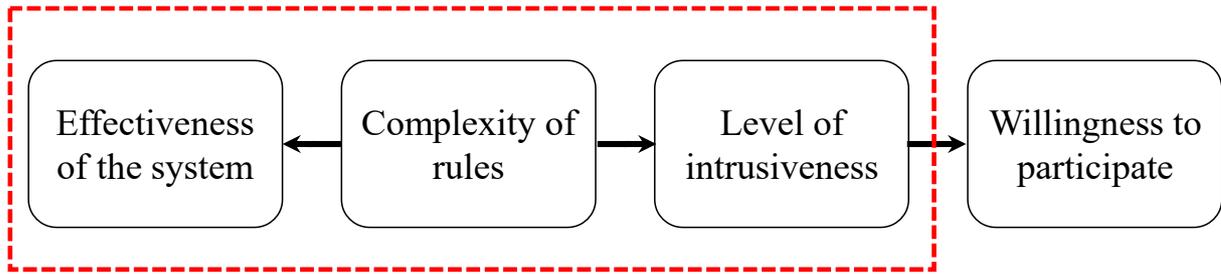


Figure 1: Relations regarding complexity of rules, level of intrusiveness and system design

2 Case study

In order to understand participants' thinking and behavior more effectively, this research uses experiments as case study. In the experiments, volunteers simulate commuting behavior on bicycles in an area in The Hague. Each volunteer follows his/her own route repeatedly. The volunteers are also asked to deliver packages of mangoes to specific locations. When they run into each other, they may pass on the packages until the mangoes arrive at the destination. The package is GPS tracked. Thus, we can observe how mango packages are transported in this area. This section starts with a brief overview of self-organizing logistics systems to motivate our design approach for the experiments.

A crowdsourced delivery system is also a self-organizing system. These bio-inspired systems, sometimes with high complexity, are based on entities that exhibit rather simple behaviors (Leitão et al. 2012, Bartholdi et al. 2010). An ant colony is a great example: each ant follows rather simple patterns of behavior, but can form a crowd that can carry out highly complex tasks. In the same way, to mimic such a self-organizing system, a well-defined set of rules is significant. Because the complexity of rules imposed on each individual is closely related to the effectiveness of the system, as well as the level of intrusiveness brought to each individual participant. A balance needs to be considered in designing a crowdsourced delivery system: the set of rules should be simple from each participant, but also able to facilitate a logistic system that is complex enough to accomplish its tasks in a practical and profitable way.

Fig.1 illustrates this trade off: more complex rules and tighter constraints lead to higher efficiency in achieving better performance at a system level; On the other hand, it may be highly disruptive, as we use the term "level of intrusiveness" to describe the tendency that following rules brings disruptions to participants daily lives. A higher level of intrusiveness will likely decrease the willingness of participants, for they need to go further to fulfill crowdsourced tasks. The experiment design in this study considers the complexity of rules and its impacts on both system level and individual level, to gain insight and give suggestions on designing crowdsourced logistics systems with a balance of system-wise effectiveness and level of intrusiveness on the participants.

We name our case study "Contingent Cycle Courier" (CCC) project. To best simulate the ad-hoc nature of the crowdsourced activities, the CCC project adopts a relay approach for parcel deliveries. In this approach, small-sized parcels are delivered to their destinations only by making use of accidental carrying power of cyclists in a city. Each parcel may "hitchhike" with several cyclists one after another, before it reaches its destination. In comparison with the traditional point-to-point (Arslan et al. 2018) and the lately discussed hub-and-spoke (Ballot et al. 2012) methods, this approach requires more collaboration among participants, and thus provides more room for them to take the initiative to make extra steps to ensure a task is successfully completed. This serves as indicators on how much intrusiveness a task brings to each individual.

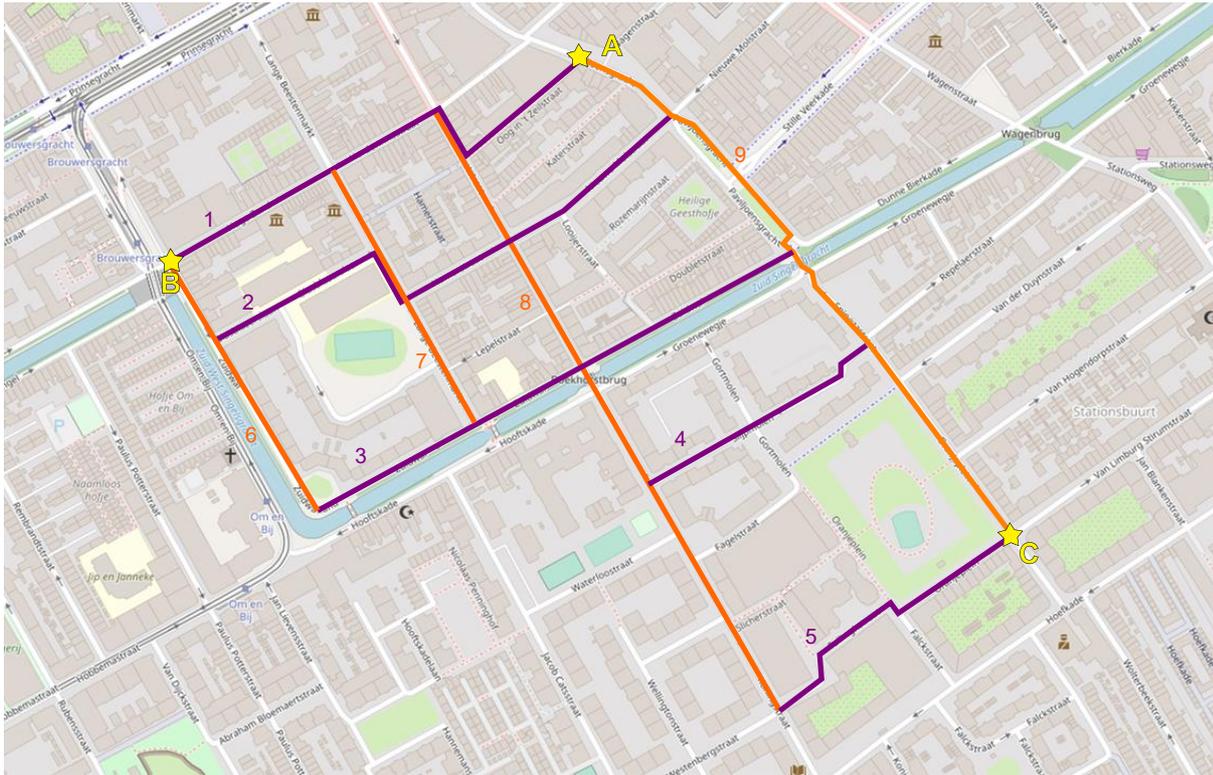


Figure 2: Selected routes numbered from 1 - 9. The map is derived from OpenStreetMap.

2.1 Route selection and parcel design

We invited 9 volunteers for parcel delivery. To simulate participants' different commuting routes, we chose an area in The Hague as shown in Fig.2. For each participant a route was selected and numbered from 1 - 9, and they traveled back-and-forth using bicycles along their designated routes. We took into consideration of the urban traffic and the safety of the participants. Some areas with complex traffic conditions were avoided. Before starting the experiments, the participants were gathered indoors to practice the activity using smaller scale simulations so that they became familiar with the rules.

For the CCC project, we designed parcels that are easy to be carried on bicycles. The small parcel was given a nick-name "Mango Equivalent Unit" (MEU). Fig.3 shows the design and actual size of an MEU.

At each of the points A, B, and C shown in Fig.2, a crew member was present to give out or to collect MEUs. Before each MEU was given out, a GPS tracker was placed, with the full awareness of each participants, to track the movements of mangoes. The tracking data is then used for analysis.

2.2 Scenario design

We designed 2 scenarios, each with a particular set of rules with different degrees of complexity. This was to observe the impact of complexity of rules on participants behaviors. Each scenario was experimented for 30 minutes.



Figure 3: Design of a "Mango Equivalent Unit" (MEU)

2.2.1 Scenario 1

In Scenario 1 the complexity of rules is lower. Cyclists follow the routes designated to them. When experiments start, parcels are handed over to Cyclist 1 and Cyclist 9 from point A. The cyclists can approach other cyclists they encounter when following within their own routes, to hand over a parcel. In the end, the parcels need to be delivered to point B or C.

2.2.2 Scenario 2

Scenario 2 has a higher degree of complexity on the set of rules in comparison with Scenario 1. In Scenario 2, MEUs are handed out at point B and point C. The ones from point B need to be delivered to point C, and the ones from point C need to be delivered to point B. On each parcel there is a sticker with an icon and a color to denote the expected destination of this parcel, so that each cyclist should only pass the parcel to the right person to be able to complete the delivery. To ensure the deliveries are fulfilled, we design a grid system and relating rules to help the cyclists fulfill their tasks.

The grid system is applied to all cyclists on all routes as shown in Fig.4. Each of the cyclists is assigned to one of the two dimensions of the grid system, represented by icons or colors. Cyclists traveling along the dimensions wear hats to indicate their directions. The two directions are noted with the hats they put on. For cyclists traveling in the east-west dimension, they put on a red hat when traveling towards east, and put on a blue hat when traveling towards west. For cyclists traveling in north-south dimension, they put on a hat with a "tin" logo when traveling towards north, and put on a hat with a "flower" logo when traveling towards south. In this way, the point B on the map is denoted by a "tin" icon and the blue color, representing the north-west corner. Similarly, the point C in the south-east corner is denoted with flower and red.

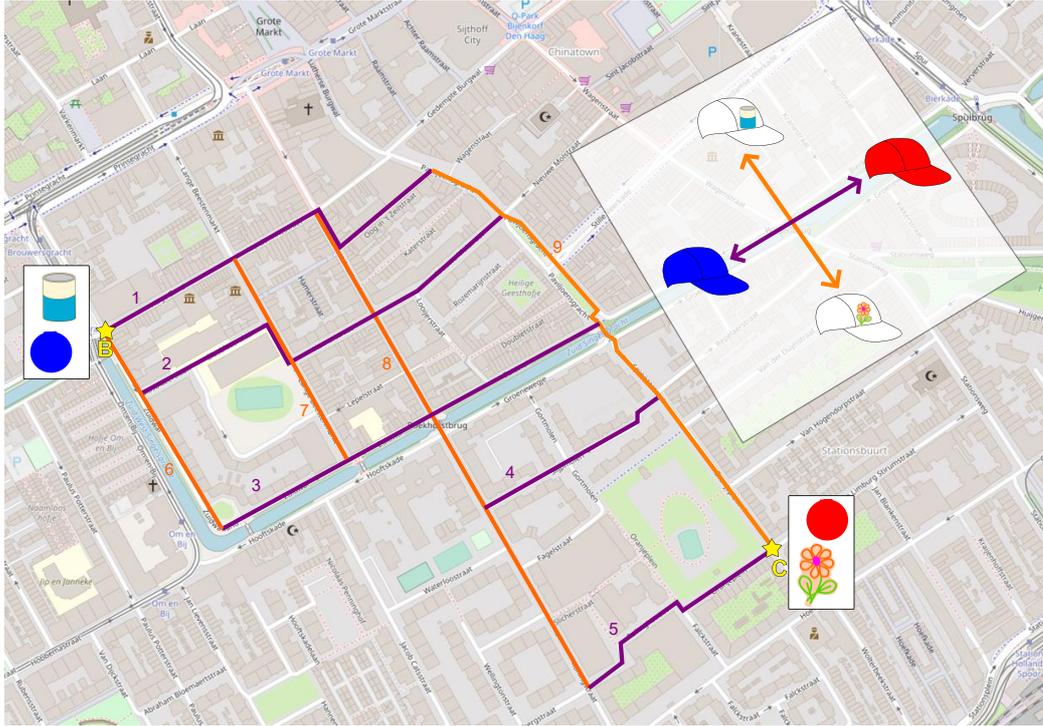


Figure 4: The routes and the grid system for delivering to specific locations. The map is derived from OpenStreetMap.

Half of the MEUs are handed out from point C, with a sticker of blue and tin denoting their destination at point B; and the other half start from point B and end at point C, which is denoted by red and flower. The cyclists carrying an MEU with the sticker blue tin, can only pass on the parcel to another cyclist with a blue hat or a hat with a tin icon. The cyclists carrying an MEU with red flower, can only pass on the parcel to another cyclist with a red hat, or a hat with a flower icon. By introducing these rules, each handing over is ensured to have the parcel one step closer to its destination.

We make a list in Tab.1 to compare the rules imposed on participants in 2 scenarios. In Scenario 1, the destination of a parcel can be Point B or C, thus only very basic rules are imposed to allow the parcels "flow" in the network. In Scenario 2, extra instructions are given to increase the efficiency of the system. Note that in both scenarios, the rules for each individual does not specify the overall objective of the system: to deliver parcels to specific points. Rather, the instructions to each individual are only to whom they can pass on the parcel. This design is in line with the principle of a self-organizing system, that simple rules imposed on each individual participant, can also achieve overall system-wise objectives that are more complex.

Table 1: Comparison of rules in Scenario 1 and Scenario 2.

	Scenario 1	Scenario 2
Follow given route	✓	✓
Receive and give out Parcels	✓	✓
Switch hats when turning around	✗	✓
Read sticker on each MEU	✗	✓
Give out parcels according to hats and stickers	✗	✓



Figure 5: The route of a mango parcel in Scenario 1

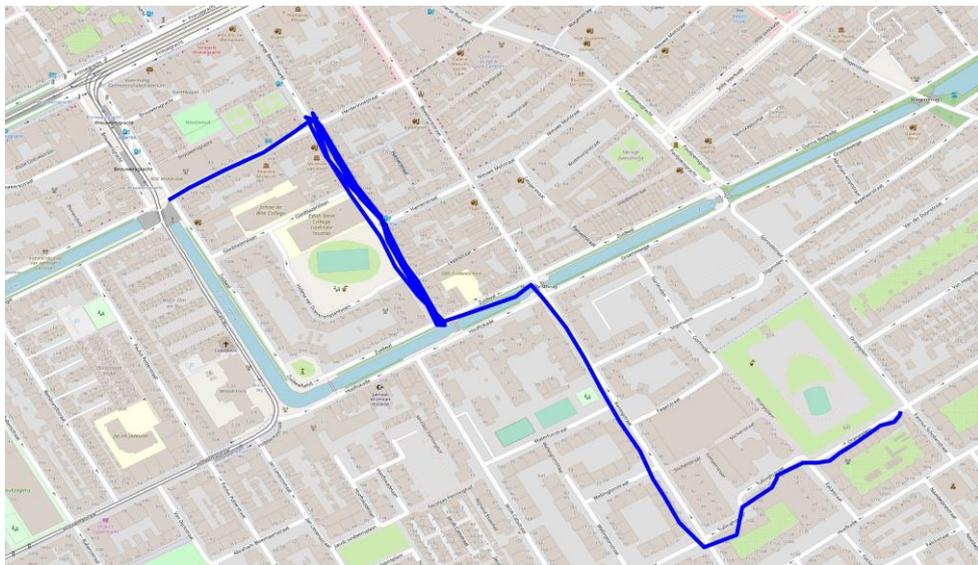


Figure 6: The route of a mango parcel in Scenario 2

3 Results and discussion

In this section, we discuss the results of the experiments by comparing the 2 scenarios. Fig.5 and Fig.6 show typical routes of a mango parcel in scenario 1 and 2, respectively. We do not directly compare the overall performance of the 2 scenarios (which are also not comparable since they have different objectives). Therefore, we choose several indicators that show the relation between the complexity of the rules, the level of intrusiveness, and how they could affect systems overall performance. This is discussed in the following parts.

3.1 Indicators

3.1.1 Pass

The number of passes denotes how many times each parcel hops from one cyclist to another. Note that when a mango carrying cyclist turns around at the end point of his journey and begins

to travel backwards (with a cap switching motion), it also counts as one pass. This indicator gives an idea how long (and how complex) the journey was, before the parcel is delivered.

3.1.2 Long-wait

Each cyclist may choose to wait at an intersection to have the parcel handed over to someone else. (This was not specified in the rule but was not forbidden either.) If a cyclist waits for more than 30 seconds at an intersection in order to give the parcel away, this pass is counted as one long waiting pass. Note that when a mango carrying cyclist turns around without giving the mango to others, it also counts as one long waiting pass if he waits for more than 30 seconds at the turning point. This indicator helps us to understand to how much extend the cyclists are willing to act in align with the rules imposed on them.

3.1.3 Turn-around

We count the number of turning-around actions of mango carrying cyclists. If a cyclist turns around with a parcel, it means the mango travels longer distance than necessary to be successfully delivered, which could potentially lead to lower efficiency of the overall system. This gives us an idea on the effectiveness of the logistics system, in particular the efficiency of relaying activities.

3.1.4 Overlap

We count the total number of routes covered by each mango for more than once. This may also help us understand how much distance each mango travels over is non-effective, which directly relates to the effectiveness of the overall logistics system.

We list the data collected from our GPS trackers in Tab.2. The list comes in two sections: total count (which includes all parcels' movements) and successful delivery (which only includes movements of parcels that are successfully delivered within the given time).

Number of successful deliveries, average number of passes, long-waits, turns, and overlaps per delivery are shown in Tab.3.

Table 2: Data collected from the GPS trackers.

		Pass	LW	Turn	Overlap
Total Count	Exp 1	54	10	30	10
			18.5%	55.6%	18.5%
	Exp 2	72	25	27	11
			34.7%	37.5%	15.3%
Successful Delivery	Exp 1	42	8	22	5
			19.0%	52.4%	11.9%
	Exp 2	54	17	19	7
			31.5%	35.2%	13.0%

Table 3: Number of deliveries, average number of passes, long waits, turns, and overlaps per delivery.

	No. Delivery	Pass	LW	Turn	Overlap
Exp 1	7	6.00	1.14	3.14	0.71
Exp 2	11	4.91	1.55	1.73	0.64

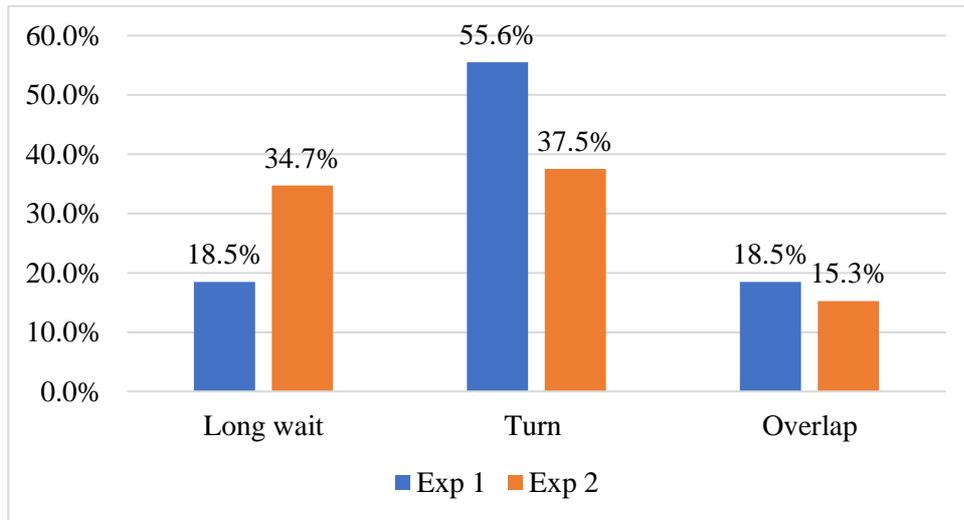


Figure 7: Long wait, turn, and overlap in percentage of all passes.

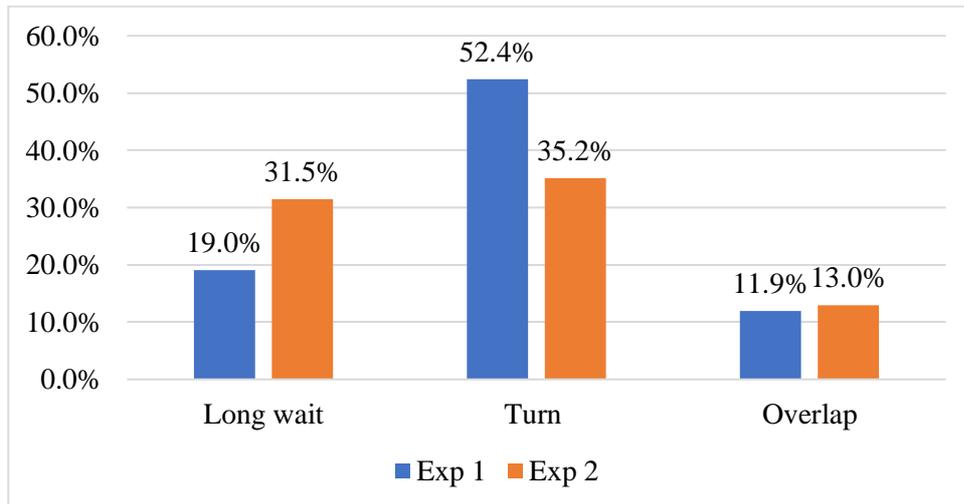


Figure 8: Long wait, turn, and overlap in percentage of passes contributing to successful deliveries.

3.2 Result findings and analysis

We do not compare the number of successful deliveries in two experiments, because this number is affected by the origin-destination arrangements. We only look into the motions of parcels to get an insight on behaviors of the participants.

More complex rules can contribute to better effectiveness. Fig.7 and Fig.8 show the percentage of long wait, turn, and overlap in all passes and in successful deliveries. In both figures, Scenario 2 has significantly lower percentage of turn-arounds among passes. This shows that less mangoes were traveling back and forth in the hands of the same participants, and that the relay of the parcels happened much faster. This is because in Scenario 2, the rules help participants identify the right person the parcel should be handed to. Thus, it decreases unnecessary trips. This shows that more complex set of rules can play a role in contributing to higher effectiveness of the whole system. This is also observed in Tab.3 in Scenario 1, each 6 passes lead to 1 successful delivery; while in Scenario 2, despite the stricter destination control rules, only 4.91 passes are needed to perform 1 successful delivery.

More complex rules bring higher intrusiveness. In both Fig.7 and Fig.8, Scenario 2 has significantly higher percentage in long waiting passes than Scenario 1 (by 87.6% and 65.8% in all passes and successful deliveries). This indicates that participants take more efforts to adjust their commuting activity in order to finish delivery tasks when rules are stricter in Scenario 2. In other words, as rules are more complex, the logistics activity becomes more demanding, and the participants react by putting more efforts to fulfill their tasks. This suggests that more complex set of rules brings higher intrusiveness to participants.

However, when rules become more complex, the increase of participants' efforts put into each *successful delivery* is less significant, as only 36.0% more long-waits per successful delivery is observed. This indicates that an increase of system performance does not necessarily require the level of intrusiveness to increase by an equivalent amount. This is especially notable for crowdsourced logistics system designers, as an effective design of the rules may increase system performance without having to raise too much the level of intrusiveness.

4 Conclusions

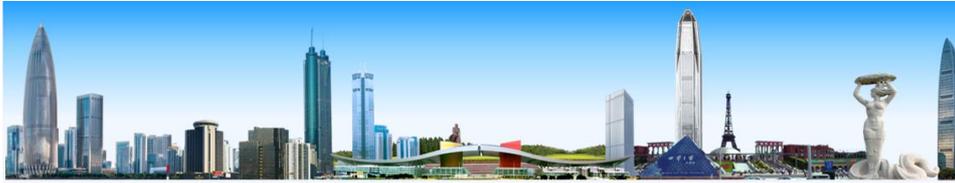
This paper uses an experimental case study to analyze the relations among system performance, complexity of rules and level of intrusiveness, in the organization of a crowdsourced logistics system. We recruited volunteers to participate in the case study, where they simulate their daily commuting actions on bicycles. In the meantime, we use their carrying capacity to move small parcels of mangoes and eventually deliver to certain locations. Each parcel was tracked by GPS trackers. The experiments were done in an area in The Hague, Netherlands. Results from the GPS were retrieved and analyzed.

From the analysis we can draw useful information regarding crowdsourced logistics systems. First, more complex rules bring higher level of intrusiveness. Thus participants, apart from their primary goals (i.e., their daily lives), may need to take extra steps, mentally and in practice, to follow the instructions given by the logistics system. Secondly, more complex rules may contribute to better overall system performance. In addition, our analysis indicates that when rules become more complex, the increase of effectiveness of the system may not be of the same amount with the increase of the level of intrusiveness. This is especially noteworthy, because it shows the significance of the design of rules of the crowdsourced systems: a well-designed system can accomplish much without being too intrusive and having to require more than necessary from participants. The study also provides reference from an economic perspective, as more intrusiveness could be paired with higher rewards, which keeps the logistics activity attractive to participants.

There are certain limitations of this study. Firstly, the experiments are only with 2 comparing groups, which makes it difficult to quantify the complexity of rules and the level of intrusiveness, which does not support deeper, more thorough quantitative studies. Subsequently, the participants in the experiments might see these tasks as their primary goal rather than the secondary, as the simulated commuting behavior is not their actual commuting behavior. Nevertheless, it does not diminish the value of this study, as it points out the importance of system design in crowdsourced logistics systems. It also reveals directions for further and more thorough study on crowdsourced logistics systems. In future research, it is worthwhile to conduct larger size experiments from people's real daily activities. It is also interesting to quantify level of intrusiveness, as in this paper, we only discuss it in a qualitative manner.

References

- A. Devari, A.G. Nikolaev, Q. He, Crowdsourcing the last mile delivery of online orders by exploiting the social networks of retail store customers, *Transportation Research Part E: Logistics and Transportation Review* 105 (2017) 105 – 122.
- H.B. Rai, S. Verlinde, J. Merckx, C. Macharis, Crowd logistics: an opportunity for more sustainable urban freight transport ?, *European Transport Research Review* 9 (3) (2017) 39.
- J.-F. Rouges, B. Montreuil, Crowdsourcing delivery: New interconnected business models to reinvent delivery, in *Proceedings of the 1st International Physical Internet Conference*, Quebec City, Canada, 2014, pp. 1 – 19.
- H. Hodson, Hand-delivered parcels find their way to you via the crowd, *New Scientist* 219 (2917) (2013) 17 – 18.
- A. Sadilek, J. Krumm, E. Horvitz, Crowdphysics: Planned and opportunistic crowdsourcing for physical tasks, in *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, Cambridge, MA, 2013, pp. 536 – 545.
- C. Chen, S.-F. Cheng, H.C. Lau, A. Misra, Towards city-scale mobile crowdsourcing: Task recommendations under trajectory uncertainties, in *Proceeding of the 24th International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, 2015, pp. 1113 – 1119.
- D. Soto Setzke, C. Pflügler, M. Schrieck, S. Fröhlich, M. Wiesche, H. Krcmar, Matching drivers and transportation requests in crowdsourced delivery systems, in *Proceedings of the 23rd Americas Conference on Information systems*, Boston, MA, 2017, pp. 1 – 10.
- C. Chen, D. Zhang, X. Ma, B. Guo, L. Wang, Y. Wang, E. Sha, Crowddeliver: planning city-wide package delivery paths leveraging the crowd of taxis, *IEEE Transactions on Intelligent Transportation Systems* 18 (6) (2016) 1478 – 1496.
- A. M. Arslan, N. Agatz, L. Kroon, R. Zuidwijk, Crowdsourced delivery – a dynamic pickup and delivery problem with ad hoc drivers, *Transportation Science* 53 (1) (2018) 222 – 235.
- Y. Kim, D. Gergle, H. Zhang, Hit-or-wait: Coordinating opportunistic low-effort contributions to achieve global outcomes in on-the-go crowdsourcing, in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montréal, QC, Canada, 2018, pp. 96.
- P.-Y. Chi, A. Batra, M. Hsu, Mobile crowdsourcing in the wild: challenges from a global community, in *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, ACM, Barcelona, Spain, 2018, pp. 410 – 415.
- L. Zheng, L. Chen, Maximizing acceptance in rejection-aware spatial crowdsourcing, *IEEE Transactions on Knowledge and Data Engineering* 29 (9) (2017) 1943 – 1956.
- J. Miller, Y. Nie, A. Stathopoulos, Crowdsourced urban package delivery: Modeling traveler willingness to work as crowdshippers, *Transportation Research Record* 2610 (1) (2017) 67 – 75.
- P. Leitão, J. Barbosa, D. Trentesaux, Bio-inspired multi-agent systems for reconfigurable manufacturing systems, *Engineering Applications of Artificial Intelligence* 25 (5) (2012) 934 – 944.
- J.J. Bartholdi III, D.D. Eisenstein, Y.F. Lim, Self-organizing logistics systems, *Annual Reviews in Control* 34 (1) (2010) 111 – 117.
- E. Ballot, O. Gobet, B. Montreuil, Physical Internet enabled open hub network design for distributed networked operations, in : *Service orientation in holonic and multi-agent manufacturing control*, Springer, 2012, pp. 279 – 292.



Place your logo here

IPIC 2020 | 7th International Physical Internet Conference | Shenzhen

Hierarchical Staffing Problem in Nursing Homes

Ting Zhang¹, Shuqing Liu^{2*}, Ping Feng¹, Yali Zheng¹, Wenge Chen²

1 Guangdong Rail Transit Intelligent Operation and Maintenance Technology Development Center, Shenzhen Technology University, Shenzhen 518118, China

2 Guangdong University of Technology, Guangzhou 510006, China

Corresponding author: liushuqing@mail2.gdut.edu.cn

Abstract: *With the increasing trend of the aging population, nursing homes have gradually become more and more important in society. Nursing work has the characteristics of "multiple shifts, high time-varying demand, hierarchical and collaborative". Multiple shifts indicates multiple shifts can cover the same time period in 24h of a day; high time-varying demand means the demand in each period may change greatly in a day; hierarchical indicates nurses have different levels; collaborative means employees with different levels cooperate to serve the elderly. This paper involves the shift design problem(SDP) and hierarchical staffing problem, and a two-stage modeling method is adopted. We will design shifts and determine the number of nurses needed for each shift in the first stage and the number of nurses with different levels for each shift is further determined in the second stage. Besides, we also made some sensitivity analysis about the impacts of different matching ratios on the total costs, which obtained some useful conclusions. The results can provide effective enlightenment and rich significance to solve practical problems.*

Keywords: *Aging population; Nursing homes; Shift design; Hierarchical staffing; Healthcare; Collaborative; Sensitivity analysis.*

1 Introduction

With the increase of the aging population in society, the number of the elderly people is also increasing fast in our country. The primary choice for the elderly is home-based care and nursing homes, but it is obvious that the home-based care can't meet the current increasing demand, and the nursing institutions have gradually developed into the rigid needs of the society. However, the workload of nursers who are specialized for the elderly is very heavy and their wages are low. Therefore, many young people, especially those with education, are not willing to work in nursing homes. As a result, it is difficult to recruit employees in nursing institutions, and there is often a shortage of staff when allocating employees. On the other hand, it's also very hard for the elderly employees to bear such a heavy workload. Therefore, the service quality will be declined due to the poor service and staff shortage, which will further reduce the number of customers and the incomes of the nursing homes. This will also lead to the reduction of wages paid to employees, and the nursing homes will fall into the dilemma of recruitment difficulties, and then form a vicious circle.

According to our investigation, the professional nurses who take care of the elderly in nursing homes also have the following characteristics: (1)the number of nurses required within 24 hours a day will fluctuate greatly according to different time periods; (2)multiple shifts can cover the same time period in a day; (3)employees will be classified according to their responsibilities and abilities, and employees with higher levels cost more; (4)due to the different levels of nurses can provide different service, we need different levels of nurses to cooperate to serve the elderly people, and there is a matching ratio between senior employees and junior employees, which means the number of junior employees that a senior employee can lead can't exceed the matching ratio. According to the above characteristics, we can find

that the nursing staff problem in nursing homes is a complex combination optimization problem. However, most of the nursing institutions are still manual scheduling at present, which not only takes more time, but also has low efficiency. Moreover, manual scheduling is easy to cause lacking of employees in peak period and surplus of employees in low peak period, which usually leads to the waste of human resources and cost. Unreasonable staffing has become a common problem in most nursing institutions, and it will aggravate the problem of recruitment difficulties in nursing homes. Therefore, an efficient and systematic method is urgently needed to solve the staffing problem in nursing homes.

Shift design problem involves in many fields, such as bank, call center, etc. Musliu et al. studied the shift design problem according to the working characteristics of call center and bank respectively. They divided a day into multiple periods according to the time period, and considered the problems of overstaffing and understaffing in each period. There are also some shift design problems including break windows, Aykin et al. designed flexible rest windows to meet employees' rest time and reduce cost. However, these shift designs seldom consider that employees can be divided into different levels, and the situation that senior employees can lead junior employees was not considered, which must be involved in our problems.

Personnel scheduling problem and task scheduling problem have similarities, and they are also very common in real situations. Seckiner et al. divided the employees into different levels, but their problems didn't consider the high time-varying demand factors, and the rest was scheduled according to the daily needs. Moreover, these problems only considered the situation that the senior employees can substitute for junior employees, but not vice versa. However, our problems should consider the cooperation of employees at different levels to work for the elderly. Noberto et al. studied the task scheduling problem, and these problems classified employees according to their abilities. But a task should be completed by employees with the same ability instead of the cooperation between employees with different abilities, which is also different from our problem.

The shift design problem was first proposed by Dantzig and was solved by set covering method, and then some researchers used the implicit modeling method to solve the problem and made comparisons with the set covering method. The results showed that the implicit modeling method was much better than the set covering method. With the complexity of the problems, more and more scholars used two-stage modeling method to solve problems. Sana et al. showed how to use a two-stage method to solve the shift design problem and task scheduling problem in detail. Two-stage modeling method can simplify the problem and presents the model more clearly. Among these papers, Lequy et al. used heuristic rules in two stages, and Pakpoom suggested to use genetic algorithm to solve personnel scheduling problem. Besides, Sana also proved the advantages of the method in terms of the quality of the solution and the calculation time.

To sum up, the nursing work in nursing homes has the characteristics of high time-varying demand, multiple shifts, hierarchical and collaborative. In this paper, we will solve the staffing problem under these characteristics, and analyze the impact of different matching ratios on the total costs. This paper will be carried out according to the following sections. Section 2 describes the problem of nursing staffing in nursing homes and puts forward some assumptions. Section 3 establishes a two-stage model to solve the problem according to the nursing characteristics, and introduces the two-stage method and the models in detail. According to different demand distribution situations, some cases about different matching ratios will be tested in section 4, and the experimental results are displayed. Some practical

management enlightenment is obtained according to the results. Finally, section 5 concludes this article and puts forward some prospects.

2 Problem definition

At present, nursing homes are facing a difficulty of recruitment due to the characteristics of nursing work, which further leads to a poor service quality and the reduction of customers, and then decrease the incomes. At the same time, manual scheduling is aggravating this problem, so the main problem that the nursing institutions need to solve is the unreasonable staffing problem. Here are some assumptions about this problem:

1. The shift must be started and ended within the specified time, and shifts have different types according to start time, such as morning shift, day shift, afternoon shift and night shift.
2. Each shift has a shift length which should be within 4h-8h, and an employee can only work one shift a day.
3. Multiple shifts are needed and can cover the same time period each day.
4. Different demands of 24 hours may be different, but the demand of each time period should be met to avoid understaffing.
5. Employees have different levels, and the service quality that a senior worker can provide is better than that of junior employees.
6. Junior employees can not be arranged if the shift only has one employee, and they must be led by the employees with the highest level. However, the number of junior workers that a senior worker can lead should be limited.
7. Since different employees have different levels and abilities, hierarchical collaboration should be considered to provide services for the elderly.
8. Different employees have different costs. The cost of an employee includes the part of shift length and the part of the level of the employee, the employee whose level is higher and shift length is longer needs more cost.

It can be seen from the above assumptions that the nursing staff problem belongs to a more complex combinatorial optimization problem. Firstly, because the nursing work has the characteristics of high time-varying demand and multiple shifts, the demand will change dramatically according to different time periods within 24 hours a day, which usually causes poor service quality due to staff shortage and waste of cost and human resources on account of overstaffing. Therefore, we need to design reasonable shifts. On the other hand, due to the characteristics of hierarchical and collaborative, the workers will be divided into several levels according to their abilities, and different levels of employees will cooperate with each other to serve the elderly, which means there have many different collocations when staffing employees. Manual scheduling is obviously inefficient. Therefore, we also need to consider the constraints of hierarchical allocation for staffing. Although these problems are very prominent in nursing homes, they still haven't found an effective way to solve them.

3 Two-stage modeling

3.1 Two-stage method

In this paper, according to the characteristics of nursing workers in nursing homes, a two-stage modeling method is proposed to solve the problem, because the two-stage model can

express the problem more clearly. On the one hand, due to the high time-varying demand and multiple shifts characteristics, it is necessary to design shifts to reduce the waste of personnel resources. In this stage, the shifts will be designed by using genetic algorithm, and the detailed time arrangement and the number of employees in each shift will be determined. On the other hand, the characteristics of hierarchical and collaborative require us to consider the cooperation and collocation of employees at all levels to reduce costs as much as possible. We will take the shift arrangement designed in the first stage as an initial condition, and further determine the final specific staffing of employees at all levels according to the constraints of hierarchical collocation by using heuristic rules. Finally, this paper also tests the influence of different matching ratios of senior employees and junior employees on the total costs under different demand distributions, and obtains some management enlightenment for practical decision-making through sensitivity analysis.

3.2 Parameter definition

For the problem in this article, we have the following definitions:

1. t stands for the start time of each shift, $t=0,1,\dots, 23$. The shift should be started within the specified time. Besides, we use h to represent the shift length, whose value is between 4 and 8. If t is 0 and h is 4, it means the shift starts at 0 o'clock in the evening with a length of 4 hours.
2. S represents each shift, and S_{th} means a shift whose start time is t and shift length is h .
3. K stands for the shift type, when k takes 1, 2, 3 and 4, it means morning shift, day shift, afternoon shift and night shift respectively. Besides, d_k means the total number of the shift type corresponding to k .
4. Each time period of a day in 24 hours were represented by i , $i=0,1,\dots, 23$. For example, when i takes 0, the time period it represents is from 0 p.m. to 1 a.m.
5. We use B to represent the personnel demand, and B_i stands for the demand in each time period, where $i=0,1,\dots, 23$, and $B_{i_{\max}}$ represents the maximum demand in 24 periods of a day.
6. m stands for the level of employees, $m=1,2, \dots, m_x$. There are m_x levels among employees in total, and the smaller the value of m , the higher the level of the employee. That means when m takes 1, it represents that the employee is in the highest level. In addition, we assume that m_a is the lowest level of the employee.
7. W represents the number of employees, W_{th} means the total number of employees in the shift whose start time is t and shift length is h , and W_{mth} means the total number of workers in the level m and in the shift whose start time is t and shift length is h .
8. C represents the cost, C_h represents the cost of a worker whose shift length is h , and C_{mh} means the cost of a worker whose level is m and the shift length is h .
9. The matching ratio of senior and junior employees is represented by n , and only the employees with the highest level can lead junior employees. Besides, the number of junior workers that a senior worker can lead is limited. For example, when n takes 3, it means a senior worker can lead at most 3 junior workers.

3.3 Models

According to the characteristics of the problem, a two-stage method is adopted and the problem is divided into two parts, which means the models are established in two stages. In the first stage, the shift design is mainly carried out to reduce the cost of human resources as much as possible under the condition of meeting each time demand. In this stage, the cost is

only related to the shift length, and the main consideration is the start time and the length of the shift. In the second stage, the shifts designed in the first stage should be taken as an initial condition, and the hierarchical collaboration should be considered on this basis to reduce the total cost as much as possible. The total cost in this stage is not only related to the shift length, but also related to the level of employees. The costs required by high-level employees and low-level employees are different, and the model is established by combining the constraints of different levels of personnel cooperation. The detailed two-stage models are as follows:

$$(IP1) \quad \min \sum_{t,h} C_h W_{th} S_{th}$$

$$\sum_{t,h} S_{th} W_{th} \geq B_i \quad (i = 0,1, \dots, 23)$$

(1)

$$S_{th} = \begin{cases} 1 & \text{when the shift } S_{th} \text{ includes the time period } i \\ 0 & \text{otherwise} \end{cases}$$

$$\sum_{t,h \in k} S_{th} \leq d_k \quad (\forall k) \quad (2)$$

$$W_{th} \leq B_{\max}(\forall t, \forall h) \quad (3)$$

$d_k, W_{th}, b_i \geq 0$ and all integer

$$(IP2) \quad \min \sum_{m,t,h} C_{mh} W_{mth}$$

$$\sum_m W_{mth} = S_{th} W_{th} (\forall S_{th} = 1) \quad (4)$$

$$W_{mth} \geq 1 (\forall m; \forall W_{th} \geq m_x) \quad (5)$$

$$W_{mth} = 0 (\forall W_{th} = 1, m = m_a) \quad (6)$$

$$W_{mth} \leq n W_{1th} (\forall W_{th} \geq 1, m = m_a) \quad (7)$$

$W_{mth} \geq 0$ and all integer

In integer programming model IP1, the objective function is to minimize the total cost of personnel resources, and the level of the employee isn't considered at this stage. While in integer programming model IP2, although the objective function also aims to minimize the total cost of human resources, different levels of employees are considered in this stage, and the total cost of an employee is also different from the first stage. Among these constraints, constraint (1) means that the arrangement of shifts must meet the need of each time period to avoid understaffing in peak period. Constraint (2) means that the number of different shift types has to be limited. Different values of k stand for different shift types, and the number of shift type which is corresponding to k can't exceed d_k . Constraint (3) shows that the total number of employees in each shift has an upper limit, which cannot exceed the maximum value in 24 periods of a day. Constraint (4) means that the total number of employees at different levels should be equal to the total number of employees on the same shift designed in the first stage. Constraint (5) indicates that there must be at least one employee in each level when the shift has enough employees. Constraint (6) indicates that junior employees

can't be arranged if the shift only has one employee. Constrains (7) means that the number of junior employees that a senior employee can lead is limited, which can't exceed n .

In the first stage, we use genetic algorithm to design shift and determine the number of employees in each shift by generating a matrix composed of the start time and the length of a shift. Each value in the matrix stands for the number of the employees in corresponding shift, and iteration is carried out to find the optimal solution. In the second stage, we use heuristic rules to determine the number of employees at different levels in each shift, such as giving priority to the lower cost collocation. For the two problems in different two stages, we use different methods to solve and iterate between the two stages to find a better solution.

4 Experiments

4.1 Parameter setting

According to the models, we set some parameters to test:

1. The shift length of one hour is 10, and then four hours cost 40 and so on.
2. The number of each shift type shouldn't exceed 3, including morning shift, day shift, afternoon shift and night shift. Table 1 shows the detailed time arrangement of the shift type.
3. Employees are divided into three levels: senior, intermediate and junior, which means the value of m_x is 3. When m takes 1, it represents the senior employee, which is also means the highest level. When m takes 3, it represents the junior employee, which means m_a is 3. The cost of a junior, intermediate and senior worker is 50, 70 and 100 respectively, and the total cost of a worker (C_{mh}) is equal to the sum of the worker's level cost and the shift length's cost.
4. We assume that the number of junior employees that a senior employee can lead could be 1, 2 or 3, which means there are 3 different situations. A senior employee can be allowed to lead at most 1, 2 or 3 junior employees respectively.
5. The demand of each time period(B_i) is irregular, and we will test the demand curves with a slow and sharp fluctuation in the case of unimodal, bimodal and trimodal respectively. In addition, we also test the case that the demand fluctuation satisfies Poisson distribution and binomial distribution, and analyze the impact of different matching ratios on the total costs under these different demand conditions.

Table 1: Time arrangement of each shift type

Shift type	Min-start	Max-start	Min-length	Max-length
Morning shift	06:00	08:00	04:00	08:00
Day shift	09:00	11:00	04:00	08:00
Afternoon shift	12:00	16:00	04:00	08:00
Night shift	22:00	00:00	04:00	08:00

4.2 Results

Since the demand of each time period must be an integer, and the characteristics of high time-varying demand may lead to different needs in different periods. In addition, the different cases of demand can be divided into unimodal, bimodal and trimodal according to the number of peaks, and the fluctuation of each peak situation can be different. We keep the maximum peak value and the minimum peak value fixed in different peak conditions, which is 30 and 5 respectively. The fluctuation situation can be divided into two types: slow and sharp. We will test the impact of different matching ratios on the total costs under different peaks and

fluctuations. At the same time, the demand curve with constant fluctuation is taken as the reference curve, and the results are compared and analyzed. In addition, Poisson distribution and binomial distribution are often used to solve the problem of meeting customer demands in reality, so we also test the influence of different matching ratios on the total costs when the demand satisfies Poisson distribution and binomial distribution.

When the fluctuation is sharp or slow, the results of the total cost affected by the matching ratios under different peak conditions are shown in Figure 1 and Figure 2 respectively.

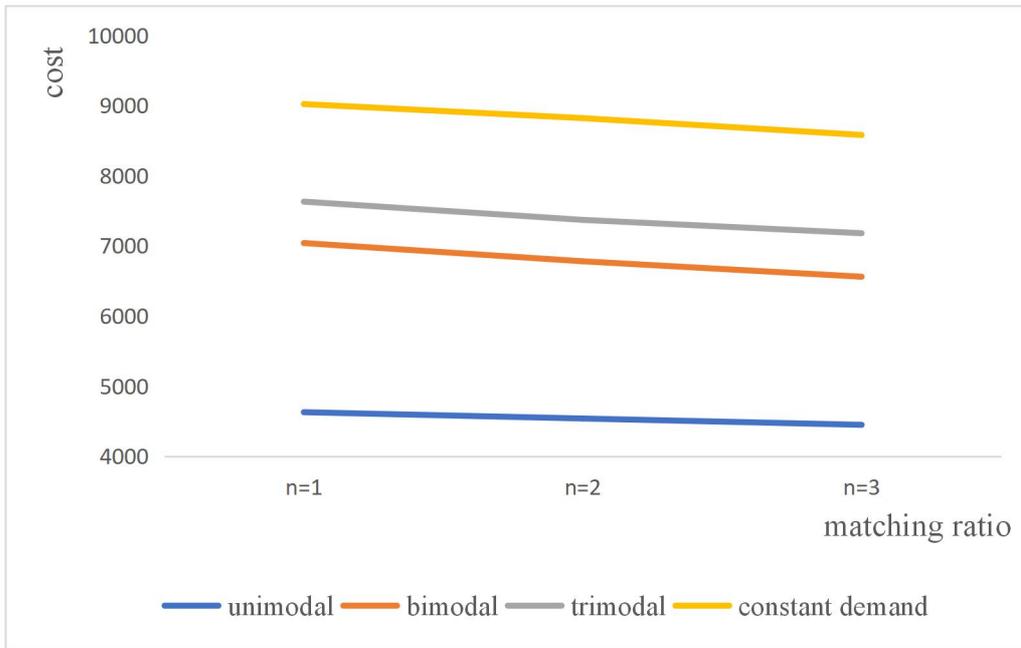


Figure 1: The influence of different matching ratios on the total costs under different peak conditions when the fluctuation is sharp

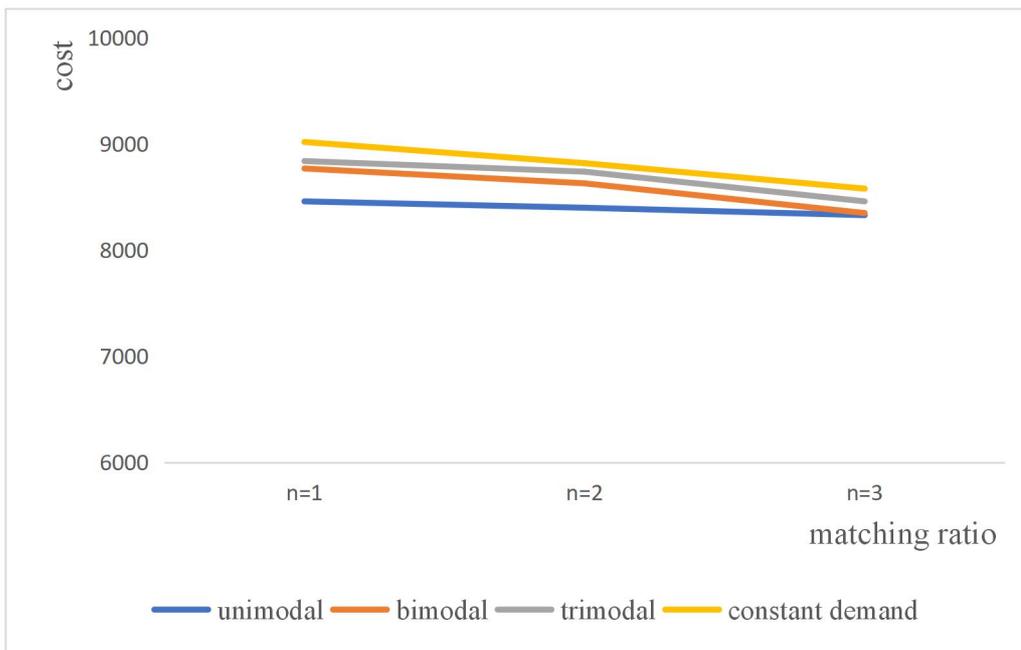


Figure 2: The influence of different matching ratios on the total costs under different peak conditions when the fluctuation is slow

According to the changes of the total costs in the above figures, the following conclusions can be drawn:

1. When the fluctuation differs greatly, the total costs of unimodal demand changes more obviously, and the growth rate of total costs of bimodal is larger than that of trimodal.
2. No matter what the fluctuation is, the total cost of unimodal demand is almost unchanged with the increase of the matching ratio. That means there is tiny differences among the total costs with the increase of matching ratio in the case of unimodal.
3. With the increase of the matching ratio, the total costs of bimodal and trimodal both show a decreasing trend. However, when the fluctuation is slow, the decreasing trend is nonlinear, and the total costs of bimodal and trimodal decrease more obviously when n increases from 2 to 3. While when the fluctuation is sharp, the decreasing trend is almost linear. The total costs of bimodal and trimodal decrease almost linearly when n increases from 1 to 3.

When the demand satisfies the Poisson distribution, we test the impact of different matching ratios on the total costs under different mean values. The results are shown in Figure 3.

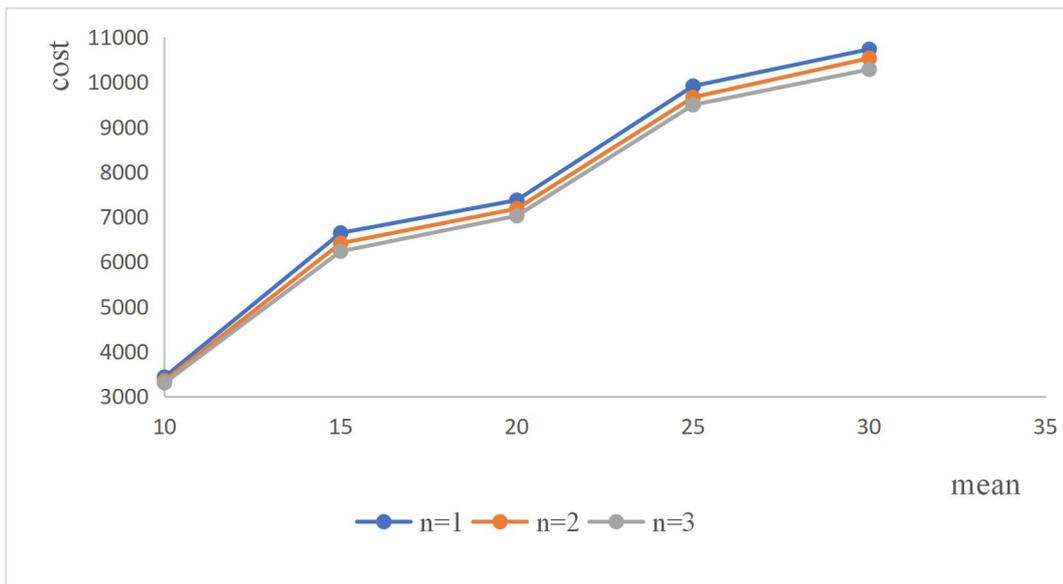


Figure 3: The influence of different matching ratios on the total costs when the demand satisfies the Poisson distribution

From this figure, we can find that in Poisson distribution, with the increase of the mean value, the total costs will gradually increase, but under the same mean value, the total cost does not differ significantly under different matching ratios. Moreover, with the increase of the mean value, the influence of different matching ratios on the total costs doesn't show an obvious change trend. In other words, even if the mean value is changed, the total cost is almost not affected by different matching ratios when the demand satisfies Poisson distribution.

When the demand satisfies the binomial distribution, we test the influence of different matching ratios on the total costs under two situations. One situation is different mean values with a fixed variance of 0.5, and the other situation is different variances with a fixed mean value of 30. The results are shown in Figure. 4 and Figure. 5 respectively.

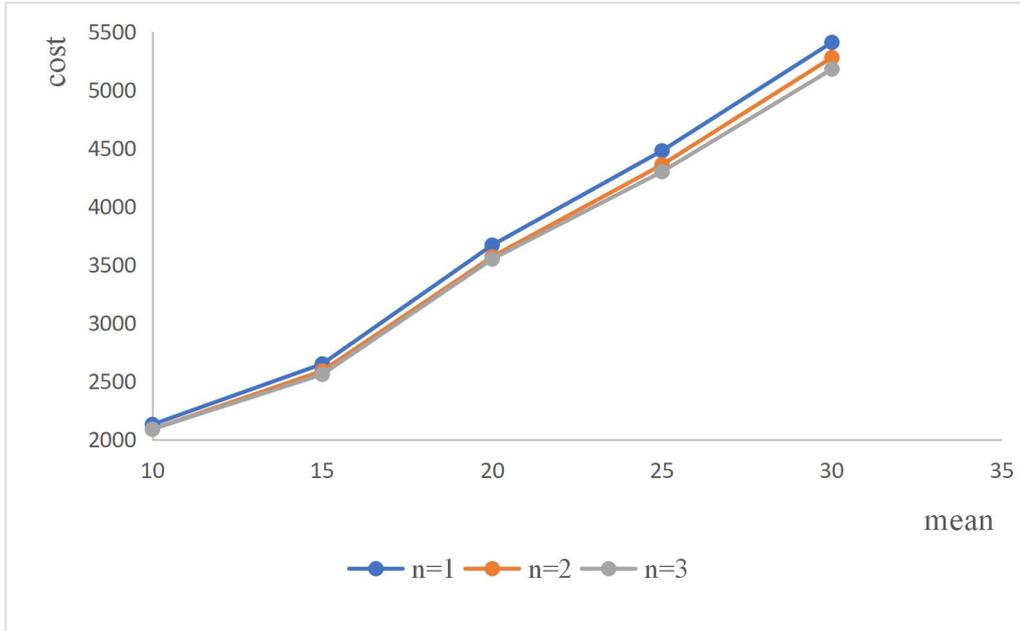


Figure 4: The influence of different matching ratios on the total costs when the demand satisfies the binomial distribution and the variance is 0.5

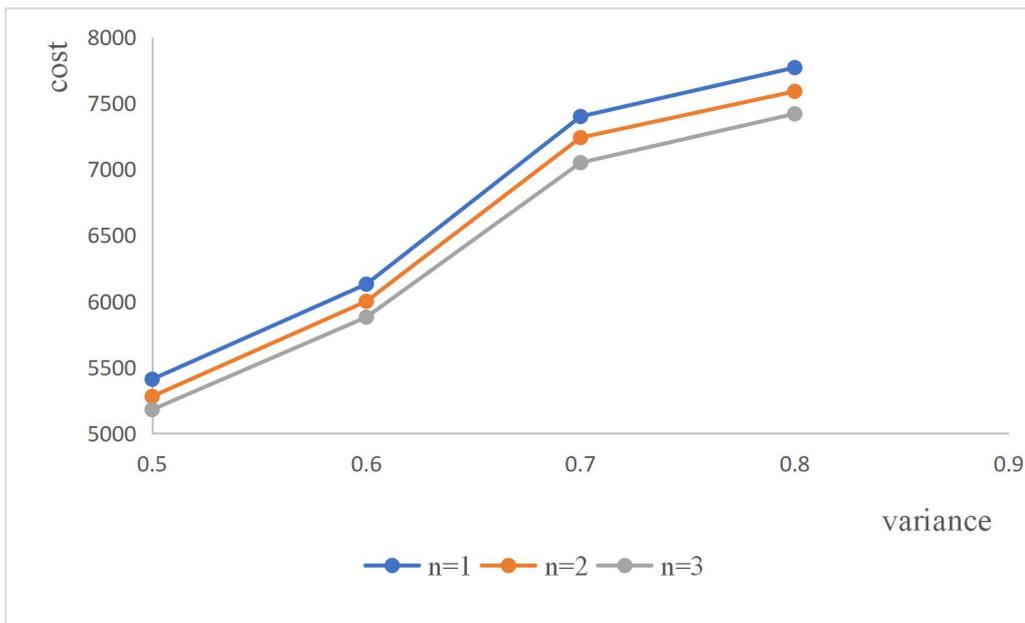


Figure 5: The influence of different matching ratios on the total costs when the demand satisfies the binomial distribution and the mean value is 30

We can draw the following conclusions according to the above figures:

1. When the demand satisfies the binomial distribution and the variance remains unchanged value of 0.5, the total cost will increase with the increase of the mean value, but the impact on the total costs is more obvious when the matching ratio is 1, and the difference is tiny when the matching ratio increases from 2 to 3.
2. When the demand satisfies the binomial distribution and the mean value remains unchanged with a value of 30, the total cost will increase when the variances increase, and

with the increase of variance, the influence of different matching ratios on the total costs becomes more and more prominently.

From the above conclusions, we can find that although the different matching ratios have a certain degree of impact on the total costs. However, when the demand situations are different, the impacts on the total costs are also different. In the case of unimodal, bimodal and trimodal, the fluctuation has a greater impact on the unimodal demand, and the bimodal demand was affected more obviously compared with the trimodal. However, when we change the matching ratio, the unimodal is the least affected by the matching ratio, while the bimodal and trimodal are greatly affected by the matching ratio, and when the fluctuation is different, the decrease trend of the bimodal and trimodal also differs. when the fluctuation is slow, the decreasing trend is nonlinear, while when the fluctuation is sharp, the decreasing trend is almost linear. In addition, when the demand satisfies the Poisson distribution, changing the matching ratio has little effect on the total costs under the same mean value. When the demand satisfies the binomial distribution with a fixed variance, the total costs are less affected by different matching ratio when changing the mean value. Only when the matching ratio is one, the total costs will differ from the other two situations prominently. On the other hand, when the demand satisfies the binomial distribution with a fixed mean value, changing the variance and the total costs are greatly affected by different matching ratios. What's more, with the increase of variance, the influences of different matching ratios on total costs are more and more significant. This shows that under the condition of binomial distribution, changing variance has a more obvious impact on the influence of different matching ratios on total costs than changing mean value. These conclusions are quite different from those summaries obtained under general conditions. We can determine which matching ratio has a greater impact on the total costs according to the distribution and fluctuation of an actual demand, which also has rich practical guiding significance for us to solve the real problems.

5 Conclusion

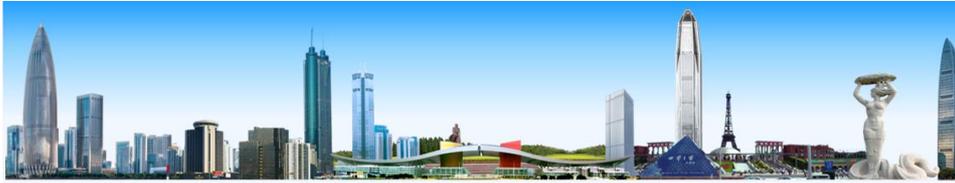
This paper mainly describes the current situation of recruitment difficulties in nursing homes, and illustrates the characteristics of nursing work and points out that the unreasonable staffing problem aggravates the recruitment difficulties in nursing homes. Therefore, we puts forward some useful methods to solve the staffing problem in nursing institutions. According to the characteristics of multiple shifts, high time-varying demand, hierarchical and collaborative, a two-stage modeling method was adopted. Genetic algorithm and heuristic rules were respectively used to resolve the problem of shift design in the first stage and hierarchical collaboration in the second stage. Finally, according to different demand fluctuations and distributions, this paper analyzes the impact of different matching ratios on the total costs and carries out a series of tests. The results show that there exists differences between the final experimental conclusions and the general conclusions, which have certain enlightenment and rich significance to solve practical problems. But in the future, we will continue to improve the model and find better methods to further study the problem.

6 Acknowledgments

This work is supported by National Natural Science Foundation of China (71701052), Guangdong Natural Science Foundation (2017C030110189), Shenzhen High-Caliber Personnel Research Start-up Project(2020111), Ordinary University Engineering Technology Development Center Project of Guangdong Province(2019GCZX006), Shenzhen Technology University Teaching Reform Project, Guangzhou Yuexiu science and technology plan project(2017-GX-005).

References

- Alex Bonutti, Sara Ceschia, Fabio De, et al.(2017): Modeling and solving a real-life multi-skill shift design problem. *Annals of Operations Research*, 252, 365-382.
- Aykin T.(2000): A comparative evaluation of modeling approaches to the labor shift scheduling problem. *European Journal of Operational Research*, 125(2), 381-397.
- Banu Sungur, Cemal Özgüven, Yasemin Kariper.(2017): Shift scheduling with break windows, ideal break periods, and ideal waiting times. *Flexible Services & Manufacturing Journal*.
- Dantzig, George B.(1954): A COMMENT ON EDIE'S "TRAFFIC DELAYS AT TOLL BOOTHS". *Journal of the Operations Research Society of America*, 2(3), 339-341.
- Hernández-Leandro Noberto A, Boyer Vincent, Salazar-Aguilar M.(2019): Angélica. A matheuristic based on Lagrangian relaxation for the multi-activity shift scheduling problem. *European Journal of Operational Research*, 859-867.
- Lequy Q , Desaulniers G, Solomon M M.(2012): A two-stage heuristic for multi-activity and task assignment to work shifts. *Computers & Industrial Engineering*, 63(4), 831-841.
- Mehran Hojati.(2018): A greedy heuristic for shift minimization personnel task scheduling problem. *Computers and Operations Research*, 66-76.
- Musliu N , Schaerf A , Slany W. (2004): Local search for shift design. *European Journal of Operational Research*, 153(1), 51-64.
- Oezgueven C, Sungur B.(2013): Integer programming models for hierarchical workforce scheduling problems including excess off-days and idle labour times. *Applied Mathematical Modelling*, 9117-9131.
- P. Pakpoom and P. Charnsethikul,(2018): A Column Generation Approach for Personnel Scheduling with Discrete Uncertain Requirements. 2nd International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Indonesia, 1-6.
- Prot,D, Lapègue,T, Bellenguez-Morineau,O.(2015): A two-phase method for the shift design and personnel task scheduling problem with equity objective. *International Journal of Production Research*, 53(24), 1-13.
- Sana Dahmen, Monia Rekik, François Soumis.(2020): A two-stage solution approach for personalized multi-department multi-day shift scheduling. *European Journal of Operational Research*, 1051-1063.
- Seckiner S U, Hadi G, Kurt M.(2007): An integer programming model for hierarchical workforce scheduling problem. *European Journal of Operational Research*, 183(2), 694-699.
- Turgut Aykin.(1996): Optimal Shift Scheduling with Multiple Break Windows. *Management Science*, vol.42, no4, 591-602.
- Volland J , Fügner, Andreas, Brunner J O.(2017): A column generation approach for the integrated shift and task scheduling problem of logistics assistants in hospitals. *European Journal of Operational Research*, 260(1), 316-334.
- Y. Chen, X. Zhang, B. Bian and H. Li.(2019): Optimal Staffing Policy in Commercial Banks Under Seasonal Demand Variation. *IEEE Access*, vol.7, 121111-121126.



Resource efficiency optimization-oriented digital twin unmanned warehouse system

Peihan Wen¹, Xuqian Ye² and Yiyang Liu³

School of Management Science and Real Estate, Jilin Yushu, China

College of Mechanical Engineering, Chongqing University, Guangdong Heyuan, China

Chongqing University-University of Cincinnati Joint Co-op Institute, Gansu Qingyang, China

Peihan Wen: wenph@cqu.edu.cn

Abstract: *With the gradual popularization of unmanned warehouse applications, resource allocation and optimization has become the key to cost control and improve customer satisfaction, especially the uncertainty of goods order arrival and the difficulty of real-time monitoring and scheduling of equipment status. In order to improve the accuracy of order forecasting and the efficiency of resource optimization, a new method of digital twin system based on an unmanned warehouse is proposed based on physical virtual fusion technology and digital twin technology. Firstly, a digital twin model suitable for a complicated warehouse system has been constructed. After that, the resource efficiency of unmanned warehouses is optimized by using the visualization technology of the model and optimization algorithm. Finally, an enterprise case is used to verify the effectiveness of this method.*

Keywords: *digital twins; unmanned warehouse; resource optimization; real-time monitoring*

1 Introduction

According to figures from the China Federation of Logistics and Purchasing (CFLP), the cost of warehousing reached 4.6 trillion yuan, which accounts for 5.1 percent of GDP in 2018. With the development of e-commerce, new retail and high-end manufacturing in China, the demand for high-standard warehousing is further increasing. Therefore, research on the problem of high cost, low service level and low efficiency is very important to improve customer satisfaction and reduce the inventory cost and the warehouse cost. In comparison with the traditional warehouse, the unmanned warehouse has the following advantages: low labor cost, minor personnel safety hazard and low risk of goods damage. So, it is very important to study the unmanned warehouse system to make the enterprise transform into information and intelligence. The unmanned warehouse system can realize the unmanned process of goods entering, tallying, storing, sorting, leaving the warehouse and order receiving.

For the problems of precise scheduling and resource optimization, the traditional solution is to consider the Problem as a Flexible Job-shop Scheduling Problem (FJSP) and solve it using immune genetic algorithms (Shi D et al. 2020), artificial bee colony algorithms (KZ Gao et al., 2016), artificial genetic algorithms (NSGA) (Ahmadi, E. et al., 2016) and other heuristic (Zhu and Zhou, 2020). By transforming the complex multi-objective optimization problem into a single-objective optimization problem (Wang et al., 2019), or treating the transportation equipment as a resource together with the three-dimensional shelf and the inspection equipment and using the genetic algorithm based on the elitist strategy to solve the problem (Ba L et al., 2016). Nevertheless, this algorithm also has some issues in solving resource allocations. The Algorithm's solution ability is limited and the degree of flexibility is not enough which cannot ensure the multi-frequency uncertain quantity goods arrived dispatch

very well. The service capacity and efficiency of warehouses is restricted by the configuration of resources such as shelves, AGVs and forklifts. As a frontier and hotspot in the field of intelligent manufacturing (Schluse M et al., 2018; Xi Vincent Wang and Linhui Wang, 2019; Tao F et al, 2017; Tao F et al, 2019), digital twin (Grieves and Vickers, 2017; Glaessgen and Stargel, 2012) is introduced into the field of unmanned warehouse in this paper. At present, there is little research on the application of digital twins in unmanned warehouses, represented by the five-dimensional model of digital twins proposed by Tao Fei et al. (2018) and its application plan in the field of warehouses. But in the manufacturing field, application is many, may take the model. Tongue X et al. (2019) applications of digital twin technology, for production and processing data is difficult for real-time interaction, based on an intelligent multi-mode Terminal Solution. Based on the digital twin 3D Model, Pai Z etc. (2020) proposed a general product-level digital twin development method in a smart manufacturing environment and validated it with 3D printing technology. Chao L et al. (2020) proposed network-based digital twin modeling and remote-control network physical system combining cyber-system and digital twin technology.

In the process of resource scheduling and efficiency optimization, the traditional genetic optimization algorithm can only optimize a single problem because of the difficulty in programming. The choice of parameters in the operator is mostly based on experience, which seriously affects the quality of the solution. Therefore, using the application of digital twin in manufacturing for reference, digital twin technology can be combined with unmanned warehouses to solve multi-resource scheduling and efficiency optimization problems. The network technology can be used to monitor the running state of the equipment dynamically and visually, which can help to deal with the problems on the spot in time and improve the operating efficiency and the ability to dispatch quickly. The parameter is set by the sensor on the spot, which does not affect the quality of the solution, so it can solve the problem of the warehouse scheduling in time and accurately, and the resource efficiency is low. Therefore, this paper carries out research in the following three aspects:

- (1)According to the process characteristics of unmanned warehouses, the architecture of digital twin unmanned warehouse systems with multi-level features is built.
- (2)The real-time mapping digital twin unmanned warehouse model is realized by using ontology modeling technology and data service systems.
- (3)Combining a neural network algorithm with clustering analysis method and genetic algorithms, an optimization analysis method for the resource efficiency of digital twin unmanned warehouses is proposed.

The rest of this work is organized as follow: The first section constructs the architecture of the digital twin unmanned warehouse system; The second section proposes the digital twin modeling scheme based on the unmanned warehouse workflow; The third section studies the analysis method of resource efficiency optimization of digital twin unmanned warehouse; The fourth section verifies the effectiveness of the above framework, model and method; The fifth section summarizes the full text and gives the future expectation.

2 Architecture of digital twin unmanned warehouse system

Based on the five-dimensional conceptual model of digital twin system proposed by Tao Fei (2017) and the characteristics of an unmanned warehouse system, a digital pairing system framework is constructed, as shown in the figure 1. The system includes: a perceivable physical layer, two technical platforms (multi-network integrated network platform and data service system platform), three layers of architecture (perception layer, a data layer and a

service layer), three databases (a local database, a system database and real-time database), five logical process mappings (inbound and outbound, tally, picking, order receiving, storage), and six types of real-time mapping entities (robots, AGV, goods, pallets, forklifts and stereoscopic warehouse).

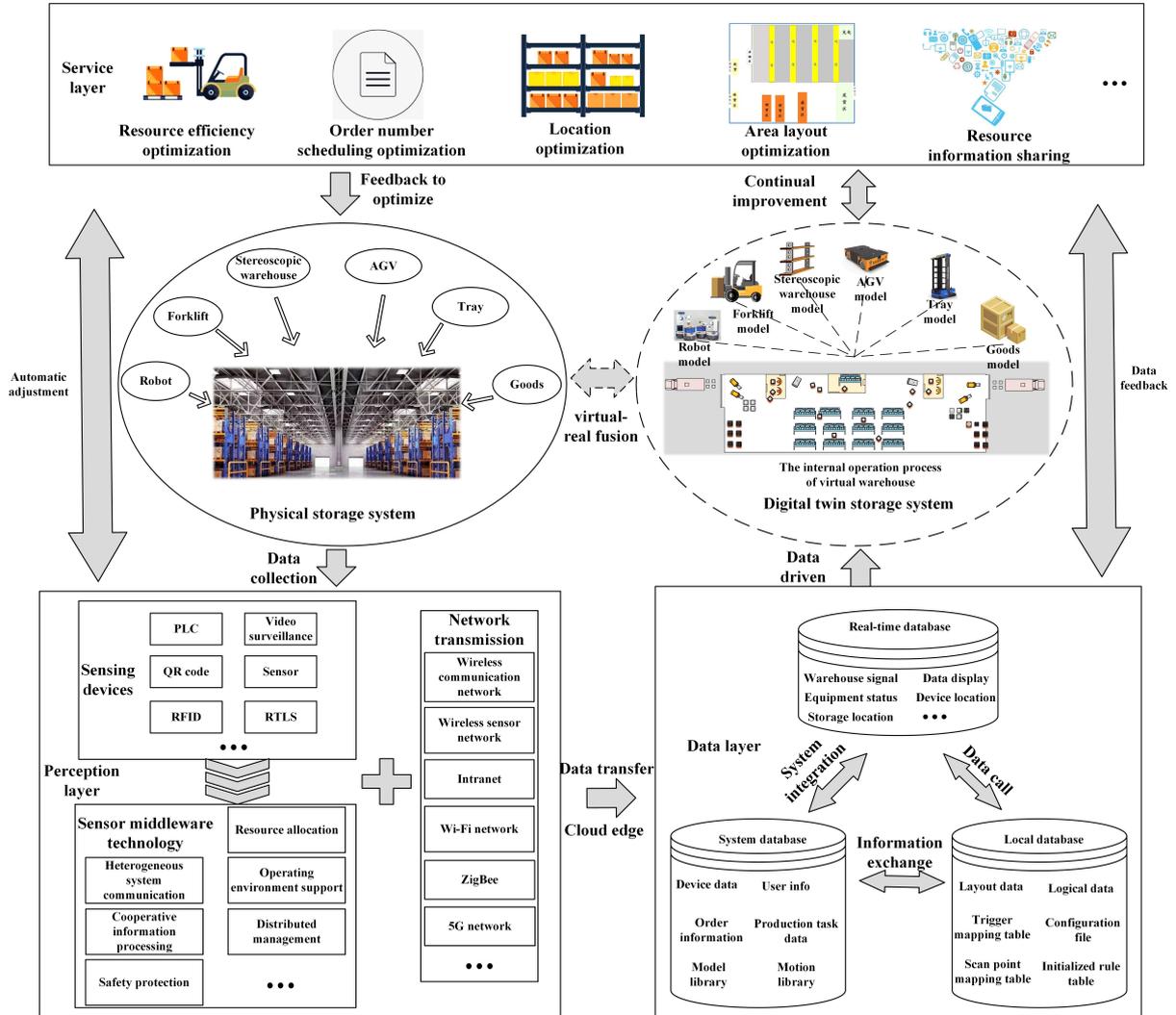


Figure 1: The frame for the digital twin unmanned warehouse system

The perception layer is used to identify objects and collect information. This layer uses sensor middleware technology to access the data which is installed on the shelves, forklifts, Automated Guided Vehicle (AGV), goods, pallets, robots, and warehouses. Then sensor middleware technology is used to process the data. Finally, the network transmission tools are applied to transmit the data to complete the collection of the bottom information.

The data layer includes: user rights management, a model interface between digital twin unmanned warehouse system and object model library, a data interface between real-time database and local database. The information contained in the real-time database, local database, system database which is shown in the frame chart: real-time database contains a warehouse signal, equipment status, a device location, a data display, a storage location; the local database contains layout data, logical data, trigger mapping table, initialized rule table, scan point mapping table and configuration file; the system database includes device data, production task data, a model library, motion library, order information and user information.

The service layer uses the drive model to run and iterate the data in the data layer to realize the functions of intelligent application, resource efficiency optimization, order scheduling optimization, storage location optimization, area optimization, resource information sharing, etc. Then the optimized information is fed back to the data center of the perception layer for virtual monitoring.

According to the frame structure of the digital twin unmanned warehouse system, the related contents in the construction of the digital twin unmanned warehouse system are further studied.

3 Model of digital twin unmanned warehouse system

The construction of digital twin unmanned warehouse models can be divided into three parts: twin modeling of entity, data service system and real-time mapping. The overall layout of the unmanned warehouse as shown in figure 2, including a three-dimensional warehouse, AGV material transportation system and other entities.

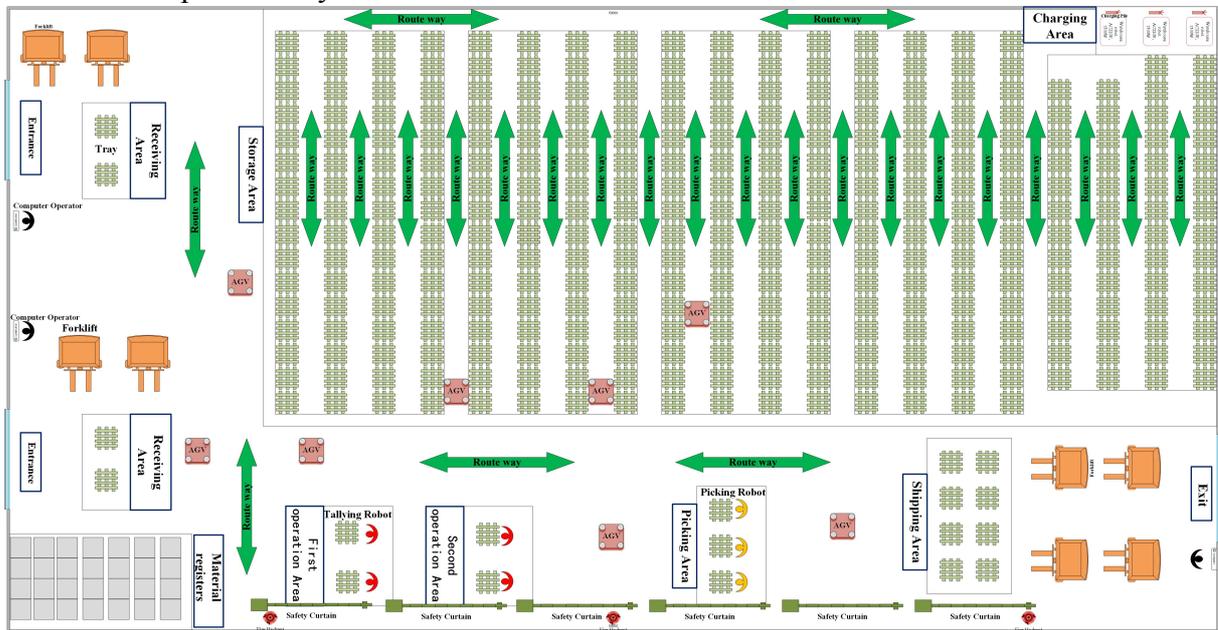


Figure 2: Unmanned warehouse layout

3.1 Entity twin modeling

The key point of ontology construction is class and attribute. Class is the definition of entity. Attribute is the description of the specific function of the class. Thus the digital twin model needs to edit the object and its attribute in advance. Then log it out and save it on the object table of the object library. Attribute, as a feature in object modeling, can be edited first. Then a multi-dimensional model of physical entity can be built using modeling software. After that the model can be imported into a simulation platform. During which the model can be lit selectively to reduce the display pressure at run time. For the moving components in the multidimensional model, they can be set as animated objects. After that, the action path of the animated components is edited. Finally, the component animation is associated with forming complete action. The attributes and interrelationships of the unmanned warehouse ontology model form a complex mesh conceptual structure as shown in figure 3.

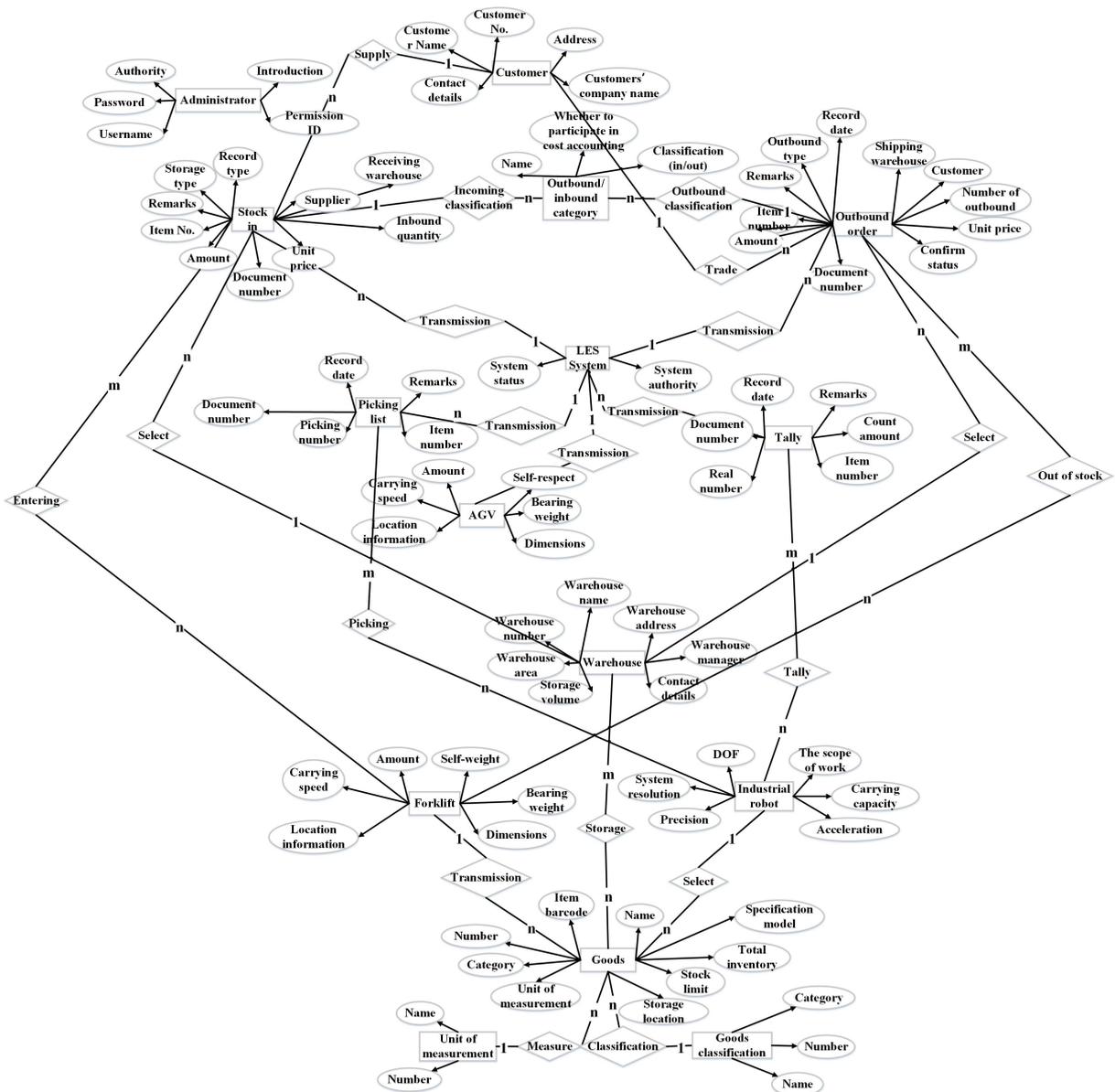


Figure 3: The mesh structure of the unmanned warehouse ontology model

The process rules and policies in the actual unmanned warehouse system are transformed into simulation logic to realize the parameterized setting of related rules and policies. The flow chart of the current unmanned warehouse is shown in figure 4.

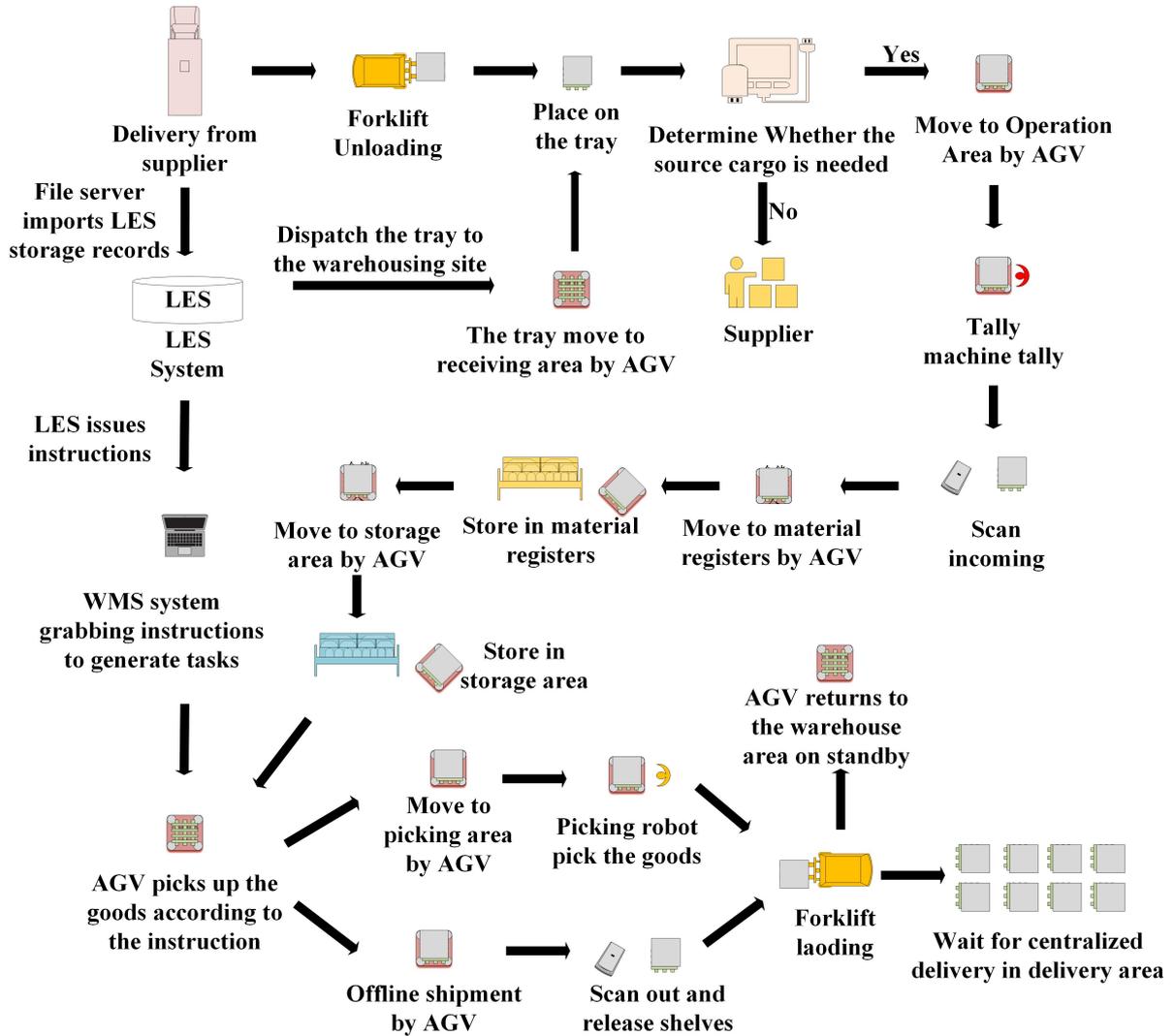


Figure 4: Flow chart of unmanned warehouse actual operation

Unmanned warehouses need to restore arriving parts and deliver them to customers. After the goods come to the supplier, the goods are unloaded by the forklift truck to the receiving area and stored on the pallets. Then the goods are transported by AGV to the tally area by the tally robot. The information is scanned to the warehouse and uploaded to the LES system. After that, the control system in the center of the unmanned warehouse will display the information about the goods' location. When the order arrives, the Les System issues instructions to the AGV vehicle. So that the AGV can go to the corresponding storage area to pick up the goods and transport them to picking area. After picking up the goods by the picking robot, the AGV is transported to the delivery area for temporary storage. Finally, when the order quantity is satisfied, the goods are transported by forklift to the shipping area for shipment.

3.2 Construction of data service system

The data service system is the connection between the real-time database, the local database and the system database. The system uses a real-time database to drive the model to run. At first, the model accesses the XML configuration file through the XML interface module to read the local database and the real-time database address information. Then the ODBC interface module is used to maintain static modeling data locally. After that, OWL-S technology is used to call the knowledge ontology of the ontology model built in owl. So that

the digital twin model can be built quickly and automatically. The regular data of the real-time database can be accessed periodically according to the time stamp through the database interface selectively. Finally, the data from the local database and the system database can be parsed. And rapid restoration of an unmanned warehouse based on a digital twin model.

3.3 Construction of real-time map

The real-time mapping rules mainly include the following:

1. Based on the actual process of unmanned warehouses, the event is coded.
2. Model initialization is to create a combination of event signals related to the object. For example, when AGV moves into a fixed position, it is necessary to record the signal identification, object identification and trigger events. It uses the servlet tool to initiate matching. During servlet initialization, the container will call the `init (servletconfig)` method, it is passed as a `servletconfig` object, called the servlet configuration object. Use this object to obtain servlet initialization parameters, Servlet name, `servletcontext` objects and so on. There are two ways to get `servletconfig` interface in a servlet:

- 1) Use `getservletconfig ()` method of the servlet, that is, `servletconfig = getservletconfig ();`
- 2) Overrides the `init (servletconfig)` method of the servlet, `public void init (servletconfig) {super. init (config); must call the init ()method of the superclass.}`

Where the Web server matches the rules of the URL:

- 1) Match the requested URL exactly to the configured URL map and call the servlet if successful, otherwise go to 2;
- 2) Try to match the longest prefix and then call the relevant servlet;
- 3) If no match can be found, use the root directory's default matching servlet or default page.

Note: When the requested URL has an extension at the end, the servlet container tries to match the servlet that handles the extension.

3. Real-time action simulation creates a trigger map in the local database, and then the device data, model data and action data in the system database are specified to realize the real-time action simulation of moving objects.

4. Real-time status display, digital twin unmanned warehouse system read data from the real-time database in a cycle according to a certain time interval to achieve workshop scene restoration and virtual transparent monitoring. On this basis, super-real-time simulation is carried out for risk assessment and optimization of unmanned warehouse scheduling. And the corresponding human-computer interface is also developed to assist decision support.

Through ontology modeling technology, the twin models have the same property and function as the physical entity in the space position, geometry size and motion characteristics. Then, the data service system is used to establish the internal and external control interfaces of the model to achieve the data interaction between the model and the three types of databases. Finally, according to the real-time mapping rule, the digital twin model can realize the effective combination of the entity elements which can complete the effective operation of the warehousing process, the tally process, the storage process, the picking process and the order receiving process, realize the whole business process of unmanned warehouse. The real-time mapping logic flow is shown in figure 5.

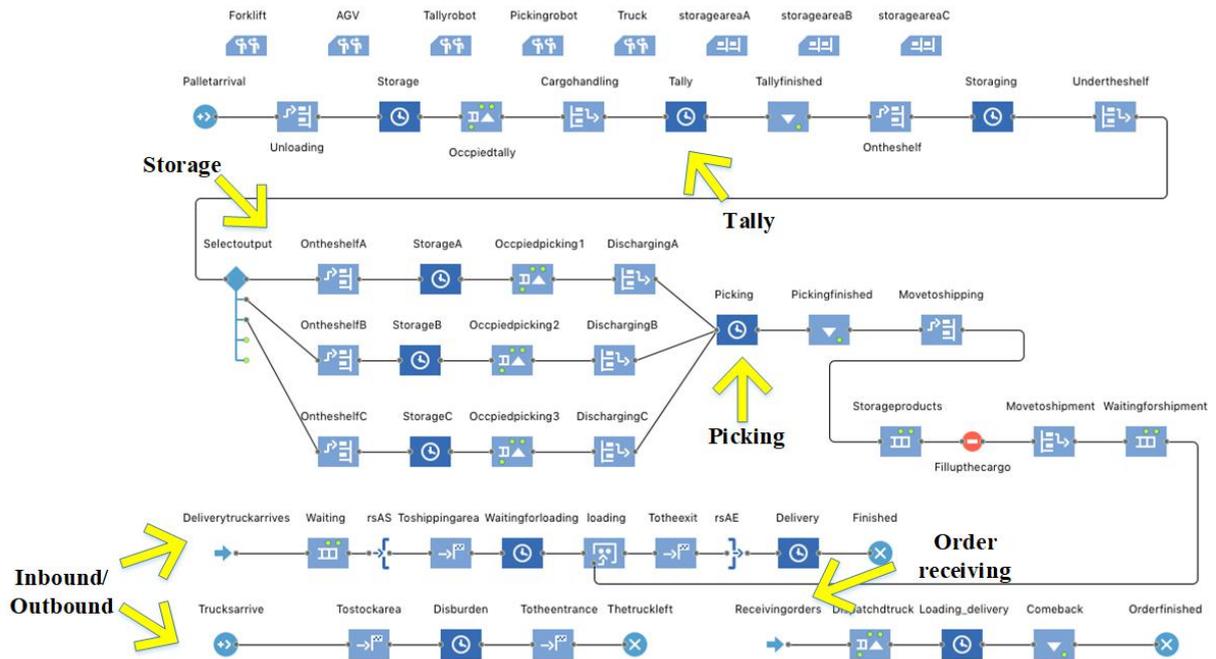


Figure 5 Digital twin unmanned warehouse real-time mapping process logic diagram

4 Optimized service of digital twin unmanned warehouse system

There are three problems in unmanned warehouses: dissatisfaction with customer order demand in time, difficulty monitoring the running status of equipment in real time and overstock increasing the cost of inventory. Traditional resource efficiency optimization methods, including simulation of the analysis and genetic algorithm optimization are based on historical data for analysis and optimization which cannot be timely feedback of real-time operation of equipment and timely processing of related issues. There exists a certain lag. The digital twin unmanned warehouse system is used to optimize the resource efficiency which is real-time and super-real-time. It can also monitor the running state of the equipment and the relevant situation that may happen in the future. So that to deal with the abnormal situation in time, it is better to raise customer satisfaction and reduce the cost of storage and inventory.

The optimization analysis flow includes: prediction analysis based on real-time data and historical data of goods and equipment running status; analysis of resource efficiency by the cluster analysis and optimization by Genetic Algorithm; the optimized resource allocation scheme is compared with the pre-optimization scheme, and the optimized data is fed back to the data service system for vector iteration and continuous optimization of the model to provide decision support for the optimization efficiency. The details are shown in figure 6.

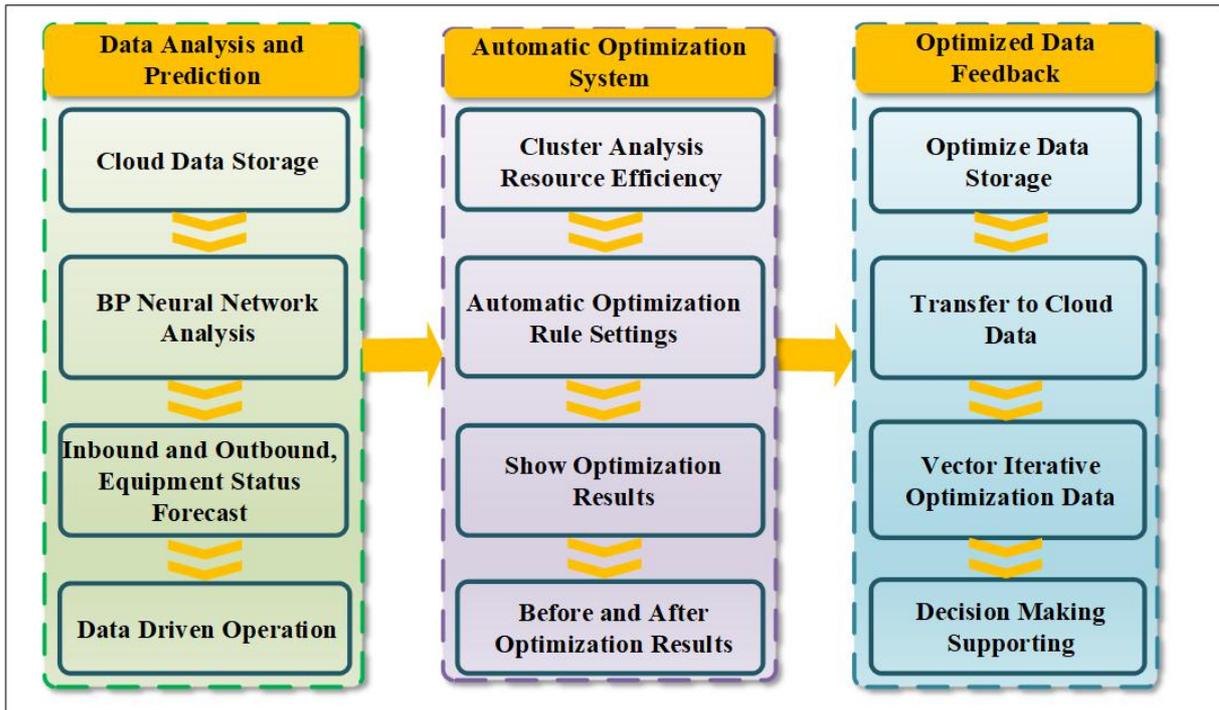


Figure 6: Resource efficiency optimization analysis process

(1) Data analysis and forecasting

The calculation is carried out by using BP neural network algorithms. The input is the relevant parameters like the historical data of the status of the goods and equipment in and out of the warehouse collected by the digital twin center, the number of hidden layers. Neural network data is divided into three parts: training data, verification data and test data. The ratio is about 7:1.5:1.5. The main aim of data training is to adjust and determine the relevant parameters of the neural network through training. The training effect of the neural network is expressed by the correlation coefficient which represents the correlation between the actual value and the predicted value. The closer to 1, the higher the fitting degree and the better the training affects prediction. The prediction data and the real-time data collected from the bottom layer are used to drive the digital twin model and feedback on the potential resource efficiency problems of unmanned warehouses.

(2) Automatic optimization of equipment resources

The process of system automatic optimization includes clustering analysis of the efficiency of the equipment and setting automatic optimization algorithms to optimize related parameters based on the analysis results.

The cluster analysis method is used to divide the resource efficiency of each kind of equipment into three categories, so as to have higher similarity within the class and lower similarity among the classes. By running the model of digital twin unmanned warehouse, the data of resource efficiency of AGV, picking robots, tally robots and forklifts are obtained. The value of resource efficiency and the direction of optimization are found by the cluster analysis. Then the genetic algorithm is used to optimize and adjust the relevant parameters to drive the model running, and realize the automatic optimization of resource efficiency.

(3) Automatic optimization of berth resources

The main objective of the optimization of shelf space is to reduce the distance of goods in the warehouse, save scheduling and transportation time. Because the structure of an unmanned warehouse is complex, it is difficult to use a specific algorithm to solve it. The system model mainly uses quantum genetic algorithms. Compared with generally used genetic algorithms, quantum genetic algorithms have better diversity and parallelism. By specifying the parameters of the shelf, inventory status and order requirements, the algorithm can be used to find the appropriate location for each batch of goods arriving. The main processes are as follows:

- 1) The population is initialized by quantum bit method according to the physical information of the shelf and the inventory state.
 - 2) The fitness of the individual is calculated by taking the shortest distance of transportation as the optimization goal and considering the principle of compatibility, the constraints of centralized stacking and the size of the shelf.
 - 3) Take the current optimal individual as the next generation's evolutionary goal. Through the quantum gate operation changes the chromosome quantum bit code, forms the next generation population.
 - 4) Repeat until the optimization criteria are met.
- (4) Optimized results feedback

The optimized resource allocation data is returned to the corresponding table of the data service system through the communication interface of the digital twin model. Then the data is transmitted back to the terminal of the service layer. Users can view the optimized model and related parameters according to the visual display interface which distributes the resources of unmanned warehouses more scientifically and reasonably.

5 Application of digital twin unmanned warehouse system

The proposed digital twin unmanned warehouse system is applied in Y company. The steps are as follows: Firstly, build the ontology model and import the data into the system database with the help of the ontology modeling software. The local database extracts the model data from PLC as the twin data of the model. Then the digital twin models are constructed by using real-time mapping technology and the running state of the unmanned warehouse is restored. Cased of this, real-time and super-real-time simulation is carried out to optimize the resource efficiency and improve the space utilization ratio of unmanned warehouses.

The data of five products in the fourth quarter of 2019 are analyzed. It was found that the number of goods in and out of the warehouse accords with Poisson Distribution. The BP neural network is used to train and forecast the order data as shown in figure 7. The correlation coefficient of the fitting degree is 0.87, which shows that the fitting degree is high and the training and forecasting effect is good. The results are uploaded to the order data table of the data service system, and the data is used to drive the digital twin model. At some time, the scene is monitoring in time and monitoring of equipment status as shown in figure 8. Including equipment operation, material storage, goods delivery, a warehouse environment, etc.

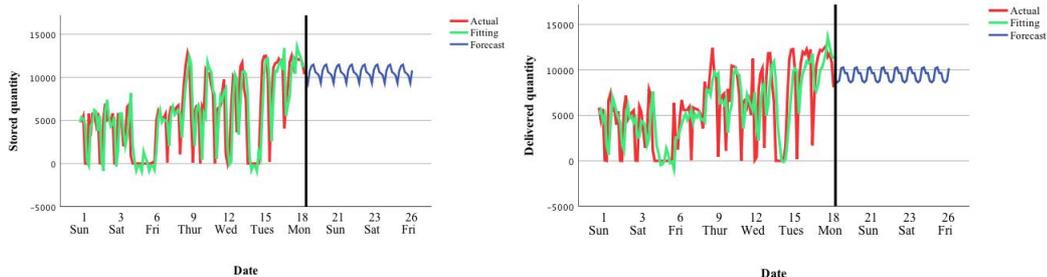


Figure 7: time series prediction of in/out storage volume



Figure 8: Large screen on site and device status data visualization

Through the data analysis of unmanned warehouse equipment resources, it can be found that the utilization ratio of automatic guided vehicles, tally robots, picking robots and other resources is unbalanced. In order to get the best allocation of resource efficiency, some rules are added to digital twin models according to the actual process and order condition. The resource quantity is adjusted dynamically by optimizing the rules.

Table 1 AGV resource utilization cluster analysis table

Auto-guided vehicles use Cluster centers				
		Club		
		1	2	3
Carriage ratio		10.83%	13.26%	21.99%
		39.57%	10.88%	27.81%
		18.18%	8.00%	40.15%
		8.42%	12.95%	36.97%
		13.01%	8.61%	8.55%

The results are as follows:

- (1) From the table, it can be seen that the carrying rate of the car is unbalanced.

(2) The data is concentrated on about 10%, and the other two cluster centers fluctuate greatly.

Therefore, the automatic adjustment of resources and working hours of AGV can make it achieve higher resource efficiency, while the same adjustment of other resources and working hours can further improve overall efficiency during the operation process.

After optimization, the resource efficiency of AGV increased from 9.12% percent to 43.86% percent, and that of picking robots increased from 15.24% percent to 39.89% percent. The other resource efficiency decreased slightly, but within the acceptable range. The turnaround time of goods was shortened from 31.29 minutes to 24.90 minutes.

Furthermore, the space utilization ratio is optimized. The shelf area of the temporary storage area and the storage area are 80 and 310 respectively. The area of the temporary storage area and the storage area are 30 and 592 respectively. In order to ensure a reasonable inventory and higher resource efficiency can reduce the number of shelves in the cargo area, the resource efficiency of the storage capacity is about 60% percent when the storage capacity reaches the peak, so it is necessary to improve the space utilization efficiency of the warehouse. When the storage resource efficiency reaches 80% percent by using quantum genetic algorithms, the number of shelves needed for the temporary storage area and the storage area are 32 and 176 respectively, and the area of the corresponding temporary storage area and the storage area are reduced accordingly. The resource efficiency of storage space has improved.

From the above analysis, we can see that the resource efficiency of the unmanned warehouse and the shelf utilization has been improved. The inventory turnover efficiency has also increased. The optimized data information is fed back to the data service system for iterative optimization and to facilitate further optimization of the model. And the data will then be fed back to the server data sharing center, scientific and rational allocation of related resources decision can be made by the decision-maker according to the visual interface model and parameters.

6 Conclusion

Digital twin technology is introduced into the unmanned warehouse system, which is helpful to monitor the running status of equipment dynamically and deal with the problems on the spot in time, to improve the running efficiency and the ability of scheduling timely of equipment and promote the realization of intelligent logistics. In this paper, a model of unmanned warehouse system based on digital twin is proposed which extends the application of digital twin technology in warehouse resource scheduling. It combines the physical entity to model ontology mapping technology and ontology translation technology, based on the data analysis and optimization technology methods (BP Neural Network Algorithm, clustering analysis method, Genetic Algorithm); the real-time data-driven digital twin unmanned warehouse system can be used for real-time transparent management and simulation decision support based on the physical information of the actual warehouse. The combination rule and dynamic priority method are designed to improve the scheduling of logistics equipment and the optimization of resource efficiency and space utilization. Digital twin modeling improves the reusability of the model, realizes the combination of virtual and real, and analyses and forecasts the future situation. The theory and method of digital twin unmanned warehouse models can be used not only for the resource efficiency and cost optimization, but also for balancing a resource efficiency and a service level. Considering the implementation cost and complexity of digital twin, there is still a big gap in the high-level integrated digital twin. Based on the model in this paper, more twin data of different objects

will be obtained and the data of different granularity will be further studied to better realize the work of "controlling reality by virtual".

Funding

This project was supported by the National Natural Science Foundation.China(No.71501020).

References

- Ahmadi, E., Zandieh, M., Farrokh, M., & Emami, S. M. (2016). A multi objective optimization approach for flexible job shop scheduling problem under random machine breakdown by evolutionary algorithms. *Computers & Operations Research*, 73, 56-66.
- Ba, L., Li, Y., Yang, M. S., Gao, X. Q., & Liu, Y. (2016) . Modelling and simulation of a multi-resource flexible job-shop scheduling. *International Journal of Simulation Modelling*, 15(1), 157-169.
- Gao, K. Z., Suganthan, P. N., Pan, Q. K., Tasgetiren, M. F., & Sadollah, A. (2016). Artificial bee colony algorithm for scheduling and rescheduling fuzzy flexible job shop problem with new job insertion. *Knowledge-based systems*, 109, 1-16.
- Glaessgen, E., & Stargel, D. (2012, April). The digital twin paradigm for future NASA and US Air Force vehicles. *In 53rd AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference 20th AIAA/ASME/AHS adaptive structures conference 14th AIAA* (p. 1818).
- Grieves, M., & Vickers, J. (2017). Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. *In Transdisciplinary perspectives on complex systems* (pp. 85-113). Springer, Cham.
- Liu, C., Jiang, P., & Jiang, W. (2020). Web-based digital twin modeling and remote control of cyber-physical production systems. *Robotics and Computer-Integrated Manufacturing*, 64, 101956.
- Schluse, M., Priggemeyer, M., Atorf, L., & Rossmann, J. (2018). Experimentable digital twins—Streamlining simulation-based systems engineering for industry 4.0. *IEEE Transactions on Industrial Informatics*, 14(4), 1722-1731.
- Shi, D. L., Zhang, B. B., & Li, Y. (2020). A MULTI-OBJECTIVE FLEXIBLE JOB-SHOP SCHEDULING MODEL BASED ON FUZZY THEORY AND IMMUNE GENETIC ALGORITHM. *International Journal of Simulation Modelling (IJSIMM)*, 19(1).
- Tao, F., Cheng, J., & Qi, Q. (2017). IIHub: An industrial Internet-of-Things hub toward smart manufacturing based on cyber-physical system. *IEEE Transactions on Industrial Informatics*, 14(5), 2271-2280.
- Tao, F., Zhang, M., Liu, Y., & Nee, A. Y. C. (2018). Digital twin driven prognostics and health management for complex equipment. *Cirp Annals*, 67(1), 169-172.
- Tao, F., Qi, Q., Wang, L., & Nee, A. Y. C. (2019). Digital twins and cyber-physical systems toward smart manufacturing and industry 4.0: correlation and comparison. *Engineering*, 5(4), 653-661.
- Tong, X., Liu, Q., Pi, S., & Xiao, Y. (2019). Real-time machining data application and service based on IMT digital twin. *Journal of Intelligent Manufacturing*, 1-20.
- Wang, W., Yang, J., Huang, L., Proverbs, D., & Wei, J. (2019). Intelligent storage location allocation with multiple objectives for flood control materials. *Water*, 11(8), 1537.
- Wang, X. V., & Wang, L. (2019). Digital twin-based WEEE recycling, recovery and remanufacturing in the background of Industry 4.0. *International Journal of Production Research*, 57(12), 3892-3902.

- Zheng, P., & Sivabalan, A. S. (2020). A generic tri-model-based approach for product-level digital twin development in a smart manufacturing environment. *Robotics and Computer-Integrated Manufacturing*, 64, 101958.
- Zhu, Z., & Zhou, X. (2020). An efficient evolutionary grey wolf optimizer for multi-objective flexible job shop scheduling problem with hierarchical job precedence constraints. *Computers & Industrial Engineering*, 140, 106280.

Application of Internet of Things into smart home scheduling

Abstract

The Internet of Things is widely used nowadays, which enabled real-time scheduling for electricity consumption task. It not only facilitates our daily life, but also provides us new thinking about how to respond to the call of “energy conservation and emission reduction” with information and high-tech. In this paper, we apply Internet of Things into the household electricity consumption task scheduling. According to the electricity usage of residents and peak-valley price, all the tasks are classified into different kinds. Based on these elements, a multi-objective optimization model is developed, with the objectives of cutting down daily electricity consumption and electricity cost, and the constraints of household electricity load and working period, to check whether “The Internet of Things” works well or not in the field of scheduling household electricity consumption task. The solution of the model shows that due to the application of The Internet of Things, we can schedule household electricity consumption task with the reduced electricity charge, and keep the satisfaction of users at a relatively high level at the same time.

Keywords: The Internet of Things, electricity consumption task scheduling, satisfaction level, peak-valley price.

I. Introduction

The application of the Internet of Things in smart homes is very common. The Internet of Things uses front-end hardware sensing devices such as temperature, humidity, light, gas and other sensors and cameras to collect product information of traditional home appliances, including location, working status, etc [1]. Through the network layer, such as Wi-Fi and Bluetooth, information is transmitted to the control terminal. Most smart homes use APP on personal handheld terminals or multi-screen controllers to achieve remote control. The mainstream smart home products that currently existed mainly include smart air conditioners, smart refrigerators, smart washing machines, smart speakers, smart curtains, smart lights, smart sockets, smart door locks, etc. In addition, although ordinary household products do not have functions such as network transmission of information, they can also be realized by using auxiliary media such as smart sockets, so that a simple smart home environment can be quickly established, which is conducive to future smart home scheduling [2,3].

It is found that there are few households research that combine the Internet and smart homes to optimize electricity charge and users’ satisfaction [4]. The in-depth

research on the Internet and smart homes are limited, and thus, this paper tries to explore and supplement this deficiency in this area.

The rest of the paper is organized as follows. In Section 2, we propose the mathematic model. The feasibility of the model is tested through a case study in Section 3. Section 4 presents the conclusion.

II. Proposed model

Parameters and variables

X_i : index of non-deferrable appliance ($i=1,2,3,4$ correspond to air purification, water heater, dishwasher, air conditioner respectively)

Y_j : index of deferrable appliance ($i=1,2$ correspond to washing machine and charging equipment)

t : time periods for a whole day ($t=1$ indicates time period 0:00-1:00)

X_i^t : binary variable indicates the operation status of non-deferrable appliance i during period t .

Y_j^t : binary variable indicates the operation status of deferrable appliance j during period t .

M : the satisfaction level of the users about electricity consumption

N : the satisfaction level of the users about electricity charge

S : the overall satisfaction level, which is expressed as the weighted average of M and N .

E_B^t : the overall electricity load before scheduling

E_A^t : the overall electricity load after scheduling

C_x : the total working time for each non-deferrable appliance for a whole day

C_y : the total working time for each deferrable appliance for a whole day

F_i : the total working times for non-deferrable appliance in a whole day

R_t : the electricity charge for time period t

Objective function

Satisfaction level of the users

First of all, the satisfaction level of the users about electricity consumption is used as an indicator to measure the change of the users' electricity using habits. When they do not make any change, the satisfaction value is one. When they need to completely adjust the original electricity usage habits, which might greatly change the amount of electricity usage in each time period, the users' satisfaction with electricity consumption will infinitely approach zero. Therefore, the expression of satisfaction

level about electricity consumption is:

$$M = 1 - \frac{\sum_{t=1}^T |E_A^t - E_B^t|}{\sum_{t=1}^T E_B^t} \quad (1)$$

Secondly, the satisfaction level about electricity charge is an indicator to measure the changes in the electricity expenses before and after the scheduling of appliance. If the user schedules and optimizes the daily electricity usage tasks, the daily electricity cost is lower than the optimized cost before the scheduling. Satisfaction level with electricity charge will be improved, the expression is:

$$N = 1 - \frac{\sum_{t=1}^T (E_A^t \times R_t)}{\sum_{t=1}^T (E_B^t \times R_t)} \quad (2)$$

The overall satisfaction level S is expressed as the weighted average of the satisfaction level about electricity consumption and the satisfaction level about electricity charge [4].

$$S = \alpha M + \beta N \quad (3)$$

$$\alpha + \beta = 1 \quad (4)$$

Thus, the first objective function is:

$$\text{Max } S \quad (5)$$

Electricity load

Besides considering the users' satisfaction level about electricity consumption, to implement peak and valley electricity prices, we need to minimize the daily electricity consumption peak and the difference between daily electricity consumption peak and valley, the second and third objective functions are:

$$\text{Min}(\text{Max } E_A^t) \quad (6)$$

$$\text{Min}(\text{Max } E_A^t - \text{Min } E_A^t) \quad (7)$$

Weighted objective function

Considering the above objective functions, we can obtain the following objective function by weighting method:

$$\text{Min}(\gamma_1 \frac{\text{Max } E_A^t}{\text{Max } E_B^t} + \gamma_2 \frac{\text{Max } E_A^t - \text{Min } E_A^t}{\text{Max } E_B^t - \text{Min } E_B^t} - \gamma_3 S) \quad (8)$$

$$\gamma_1 + \gamma_2 + \gamma_3 = 1 \quad (9)$$

Among them, $\frac{\text{Max } E_A^t}{\text{Max } E_B^t}$ and $\frac{\text{Max } E_A^t - \text{Min } E_A^t}{\text{Max } E_B^t - \text{Min } E_B^t}$ are designed to set the target value at

around 1, which facilitates comparison with users' satisfaction level at the close range, and thus, reduces the impact of the large numerical gap.

Constraints

First of all, because the optimization is affected by the users' satisfaction level, if the daily electricity consumption after the scheduling and the user's original consumption are quite different, it will significantly change real life of the users, which is impractical. Therefore, the following constraint ensures that the user's total daily electricity consumption remains the same after re-scheduling:

$$\sum_{t=1}^T E_A^t = \sum_{t=1}^T E_B^t \quad (10)$$

Second, since electricity billing is one of the most important factors, it is necessary to ensure that the user's electricity charge is lower than the original charge after scheduling:

$$\sum_{t=1}^T (E_A^t \times R_t) \leq \sum_{t=1}^T (E_B^t \times R_t) \quad (11)$$

In addition, taking into account the threshold of the household users' electricity load, if the power consumption is too large in a short period of time, it will lead to overload protection of the electric gate, causing tripping phenomenon. Thus, after scheduling optimization, the electricity consumption at various times of the day should also be lower than the electricity gate overload protection. In this paper, the maximum carrying load of 8.8kw (=40A x 220V) of the general household electricity meter with specification 5 (40)A is used to calculate the upper limit of each household's electricity load.

Since this research excludes the scheduling of basic electricity tasks, including: refrigerator, lighting equipment, rice cooker, induction cooker, television set, vacuum cleaner. The corresponding power rates are: 0.04kw/h, 0.2kw/h, 0.5kw/h, 2kw/h, 0.1kw/h, 1.4kw/h. Therefore, after excluding the basic power tasks electricity load, the maximum load can be set at 5.5kw for each time period.

$$\forall E_A^t \leq 5.5 \quad (12)$$

III. Case study

Due to the different living habits of each family, there are some differences in the using time periods of the appliance. For example, the fifth household chooses air conditioner running time in the evening peak period of electricity consumption, and the sixth household has no special requirements on the running time of the air conditioner. The following table records the electricity consumption habits of the ten households for six appliance.

Table 1. Electricity consumption habits of ten households (hr.)

Household Appliance	1	2	3	4	5	6	7	8	9	10
Air purification	6	5	6	6	6	4	4	5	5	6
Water heater	3	2	3	3	3.5	2.5	2.5	3	2.5	3
Dishwasher	3	3	3	3	3.5	3.5	2.5	3	3	2
Washing machine	2.5	2	1.5	2.5	3	1.5	1.5	2.5	2.5	2
Air conditioner	8	7	9	8.5	4	7.5	8	8.5	8	8
Charging equipment	4.5	5	3.5	3.5	4	3.5	3.5	4	3.5	4.5

After scheduling, take first household as an example, the optimal running time for the appliance is listed as follows:

Table 2. Running time for household 1 after scheduling

Appliance	Running time after scheduling	Total running hours (hr.)
Air purification	8:00-14:00	6
Water heater	11:00-14:00	3
Dishwasher	9:00-12:00	3
Washing machine	9:00-11:30	2.5
Air conditioner	0:00-8:00	8
Charging equipment	1:00-3:00,4:00-5:00, 6:00-6:30,8:00-9:00	4.5

The following figure shows the electricity consumption before and after scheduling.

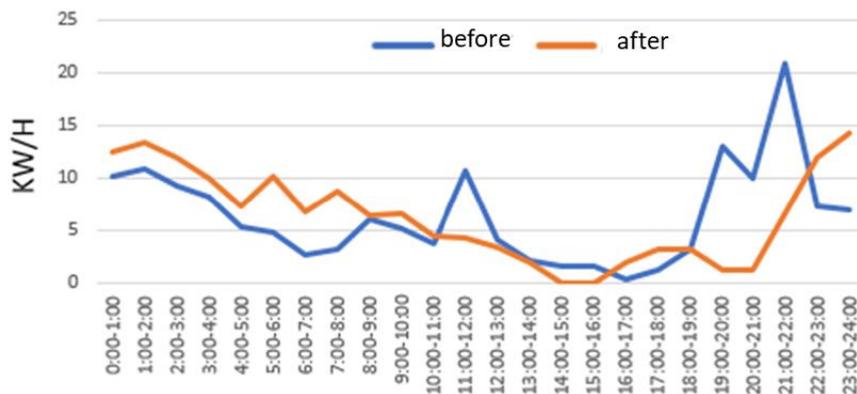


Fig. 1. Electricity consumption before and after scheduling

Through the comparison between before and after the rescheduling of daily electricity consumption of the ten households, it can be observed that the distribution of daily electricity consumption of residents has changed significantly, and the trend of electricity consumption throughout the day has fluctuated and decreased from 0:00 onwards, and goes up slightly in the evening. For example, during the original period

of 19:00-23:00, when residents used flexible power tasks, and under the effect of scheduling optimization, the power consumption of that period was significantly reduced and the peak was reduced; Electricity consumption remains low from 00-17:00, while new peaks are delayed to 0:00-1:00 a.m., and electricity consumption during peak is lower than the original one. Because the start-up modes of appliance are changed from the original manual mode to the remote control, many power tasks can be scheduled during the time that the residents are not at home. Thus, the fluctuations in daily electricity consumption can also be reduced. It is explained that after optimizing the electricity task scheduling model based on the satisfaction of the user's electricity price, the difference between the peak and valley of the daily electricity consumption of these ten households has indeed been reduced.

IV. Conclusion

In this paper, we apply Internet of Things into smart home electricity consumption scheduling. This paper simultaneously takes residents' original electricity consumption habits and reducing electricity charge into consideration, combined with the current peak-to-valley electricity price policies, to provide household electricity consumption scheduling. The specific scheduling scheme was studied and the following conclusions were obtained:

After optimizing the scheduling and optimization of the collected electricity consumption of ten households, the residents' daily electricity consumption has been reduced, and the optimization effect of the situation is obvious. Residents use electricity to cut peaks and fill valleys to a certain extent;

Since the scheduling is based on the actual electricity consumption habits of residents, the optimization results firstly consider the users' experience of household users. Meanwhile, the electricity charge of each household after optimization has been reduced by 20%-30%. That is to say, under the current peak and valley electricity price policy, the Internet of Things technology can effectively solve the problem of power dispatch, while ensuring a high level of residential power satisfaction, and helping users develop an orderly and reasonable electricity consumption habit.

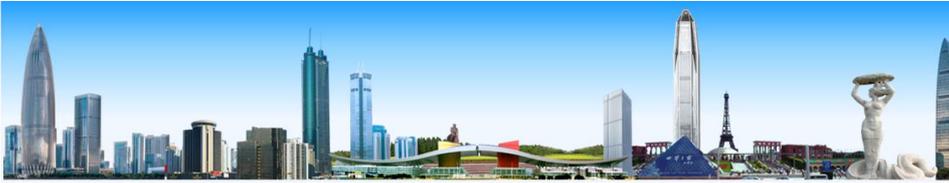
Acknowledgment

This research is financially supported by Macau University of Science and Technology (Grant No. FRG-18-022-MSB).

Reference

- [1] Erol-Kantarci M, Mouftah H T. Wireless Sensor Networks for Cost-Efficient Residential Energy Management in the Smart Grid[J]. IEEE Transactions on Smart Grid, 2011,2(2):314-325.

- [2] Du P, Lu N, Appliance commitment for Household Load Scheduling[J]. IEEE Transaction on Smart Grid, 2011, 2(2): 411-419.
- [3] S. Tang, V. Kalavally, K. Ng, et al. Development of a prototype smart home intelligent lighting control architecture using sensors onboard a mobile computing system[J]. Energy & Buildings, 2017, 138(27):15-17.
- [4] Y. Huang, K. Wang , K. Gao, T. Qu, H. Liu. 2019. Jointly Optimizing Microgrid Configuration and Energy Consumption Scheduling of Smart Homes. Swarm and Evolutionary Computation. 48, 251-261.



Integrated Production and Maintenance Scheduling using Memetic algorithm under Time-of-use Electricity tariffs

NING Tong and CHEN Jian*

Nanjing University of Aeronautics and Astronautics, Nanjing, China

Corresponding author: justinchenjian@gmail.com

Abstract: *In response to the actual needs of manufacturing enterprises for energy reduction, energy-efficient scheduling has received more and more attentions. Most research on energy-efficient scheduling in literature only consider scheduling jobs on machines. This paper investigates an integrated production and maintenance scheduling problem under time-of-use electricity tariffs to minimize total energy consumption. We consider the flexible periodic maintenance strategy and model the integrated scheduling problem as a mixed integer program model. A memetic algorithm is proposed based on genetic algorithm with a parallel local search structure enabled by simulated annealing and tabu search. The parallel local search structure achieves high-quality exploitation ability in jumping out local optimum. A two-segment coding operator is proposed taking into account both job and maintenance scheduling decisions. Besides, a novel stacking adjustment strategy is presented to improve the search efficiency of the MA. By comparing to the optimal solutions by CPLEX, the performance of the MA is verified.*

Keywords: *Time-of-use electricity tariffs, memetic algorithm, production scheduling, preventive maintenance*

1 Introduction

Energy shortages have become a bottleneck restricting the economic development of many countries. Manufacturing industry is facing unprecedented resource and environmental pressure due to its energy consumption characteristics such as high energy consumption and low energy efficiency. In recent years, the concept of Green manufacturing (GM) has gradually emerged and become an important paradigm in the transformation and upgrading of manufacturing.

The time-of-use electricity tariffs (TOU) is a demand-side power load management method that guides users to achieve a balanced output of power loads by formulating different power price ranges. In this mechanism, there are multiple electricity tariffs slots with different unit electricity prices in one day. Generally, the electricity tariffs during peak periods can reach two to three times compared to that during low peak periods. Electricity suppliers use the TOU to stimulate consumers to adjust their peak hours of electricity demand to other periods with lower prices, thereby reducing the peak-to-average ratio (PAR) of the demand load curve.

Under the TOU, how a company reasonably schedules production tasks have a huge impact on the company's power consumption. In the context of time-of-use electricity tariffs, (Moon & Park, 2014) studied the mixed processing shop scheduling problem under the TOU, and established two discrete-time mathematical models to minimize the weighted sum of the maximum completion time and the total electricity cost. (Luo et al., 2013) proposed a multi-objective ant colony optimization meta-heuristic algorithm for the mixed flow shop scheduling problem of TOU to minimize the weighted sum of total completion time and total electricity cost. (Sharma et al., 2015) studied the variable processing speed scheduling

problem under the TOU and used a multi-objective meta-heuristic optimization algorithm to minimize the total electricity cost and environmental impact. (Fang et al., 2013) studied the single-machine scheduling problem considering the TOU under the assumptions of constant processing speed and variable processing speed, and proved that the problem is a strong NP-hard problem. (Che et al., 2016) investigated a single machine scheduling problem under TOU tariffs to minimize the total electricity cost. A new continuous-time MILP model was developed. Based on the property analysis of the problem, an efficient greedy insertion heuristic was proposed. (H. Zhang et al., 2014) studied the flow shop scheduling problem under the time-of-use electricity price model, and established a discrete-time integer programming model to simultaneously minimize the total electricity cost and carbon dioxide emissions.

Obviously, most of the energy-saving scheduling research under the TOU does not consider machine failures and machine maintenance. However, machine failures and machine maintenance are common in actual production and are closely related to production scheduling. In order to reduce the impact of machine failures on production, most companies will adopt preventive maintenance strategies (Allaoui et al., 2008). Preventive maintenance aims to prevent failures. Through the inspection and detection of equipment, the signs of failure are discovered in advance, and preventive equipment maintenance is performed to keep the equipment in the specified functional state and avoid shutdown maintenance that seriously affects production.

Therefore, this paper integrates production scheduling and maintenance scheduling. For the single machine-type shop, under the TOU, we adopt the flexible periodic maintenance strategy, and use the delivery date as a hard constraint to study the integrated scheduling problem of production and machine maintenance to minimize the objective of power consumption.

First, we establish a mixed integer program (MIP) model for the integrated scheduling problem. Considering the complexity of the problem, we propose a memetic algorithm (MA) based on the population-based search of genetic algorithm (GA). We propose a two-segment coding operator to realize the integrated decision-making of job sequence and processing interval. A parallel local search enabled by simulated annealing (SA) and tabu search (TS) is embedded in the GA, which improves the algorithm's exploration and exploitation capabilities (Q. Zhang et al., 2012).

2 An integrated Model for production and maintenance scheduling under TOU

2.1 Problem description

Assume that an order has n independent jobs, the processing time of job i is t_i , and the power consumption per unit time of job i is c_i . The due date of the order is d_{\max} , that is, the completion time of all jobs of the order must be less than d_{\max} .

All the jobs are processed in a single machine. In order to avoid accidental failures of machine, a strategy of flexible periodic maintenance is adopted. The flexible periodic maintenance strategy stipulates the shortest continuous running time and the longest continuous running time of the machine; that is, the time interval between two consecutive maintenance must be not less than v and not greater than u .

Consider order production under the time-of-use electricity tariffs mechanism. This paper adopts the piecewise TOU pricing mechanism. For example, the electricity tariff cycle is 24

hours, and there are 4 electricity tariffs intervals in the cycle, and the length of each interval may not be equal. Based on this mechanism, jobs arrangement in different electricity tariffs intervals have different processing costs. Assume that the start time of the k electricity tariffs interval is b_k , the end time is b_{k+1} , and the corresponding electricity tariffs is p_k . If the processing time of the job in the k electricity tariffs interval is t_i , the power cost of the job is $t_i p_k$.

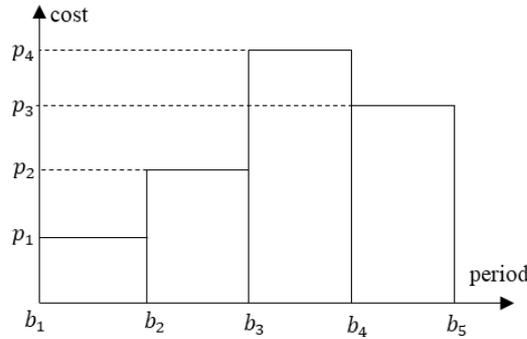


Figure 1: Schematic diagram of TOU mechanism.

In addition, the problem also satisfies the following assumptions.

- (1) Jobs processing and machine maintenance cannot be interrupted.
- (2) The machine can process at most one job at any time.
- (3) The job is allowed to be produced across the electricity tariffs interval, but it must be processed continuously without interruption.
- (4) The job processing starts from time zero.
- (5) Maintenance work can be carried out across intervals. And every maintenance time is less than the length of any electricity tariffs interval.

The goal of the integrated energy-efficient scheduling for production and maintenance is determine the processing sequence of the jobs and its start and end time, and formulate a flexible maintenance plan for the machine (the number of machine maintenance and its start and end time) to minimize the total electricity cost (TEC) of order processing under the TOU mechanism.

In order to coordinate the scheduling of order jobs and machine maintenance, this paper regards the flexible periodic maintenance of the machine as special jobs. The processing time of the job is t_i . which has processing time is t_0 but does not consume power. Due to the flexible maintenance strategy, the time interval between two consecutive maintenance must be not less than v and not greater than u . In order to reduce the impact of switch on and off on production efficiency, the number of maintenances should be minimized under the premise of ensuring the flexible periodical maintenance strategy of the machine. according the order lead time d_{\max} can be calculated that the scheduling plan requires at least maintenance times. We set the number of maintenances is m .

$$\begin{cases} m \geq \lceil d_{\max} / u \rceil \\ mt_0 + \sum_{i=1}^n t_i \leq d_{\max} \end{cases} \quad (1)$$

2.2 Modelling

Parameter:

d_{\max} : Order lead time

n : Number of order artifacts

m : Number of maintenance parts

N : Total number of jobs (total number of order jobs and maintenance jobs)

M : The number of electricity tariffs intervals.

t_i : The processing time of the job.

c_i : The unit power consumption rate of the job.

$c_i = 0; n < i \leq N$: The unit power consumption rate of the maintenance job.

$t_i = t_0; n < i \leq N$: Time to maintain artifacts.

v : The shortest time interval between two maintenance.

u : The longest time interval between two maintenance.

b_k : The start time of the first electricity tariffs interval.

b_{k+1} : End time of the first electricity tariffs interval.

p_k : The electricity tariffs in the k -th electricity tariffs interval.

A : Infinite number.

Decision variables:

$x_{i,k}$: The processing time of the job in the electricity tariffs range, $1 \leq k \leq M, 1 \leq i \leq N$

$y_{i,k}$: 0-1 variable, if the job is processed in the electricity tariffs range, its value is 1, otherwise it is 0. $1 \leq k \leq M, 1 \leq i \leq N$

The MIP model:

$$\min TEC = \sum_{i=1}^N \sum_{k=1}^M p_k c_i x_{i,k} \quad (2)$$

$$\sum_{k=1}^M x_{i,k} = t_i; 1 \leq i \leq N \quad (3)$$

$$x_{i,k} \leq t_i y_{i,k}; 1 \leq i \leq N, 1 \leq k \leq M \quad (4)$$

$$\sum_{i=1}^N x_{i,k} \leq b_{k+1} - b_k; 1 \leq k \leq M \quad (5)$$

$$y_{i,k} + y_{i,k+1} + y_{j,k} + y_{j,k+1} \leq 3; 1 \leq i \leq N, 1 \leq j \leq N, 1 \leq k \leq M - 1, i \neq j \quad (6)$$

$$\sum_{k=h+1}^{l-1} y_{i,k} \geq (l-h-1)(y_{i,l} + y_{i,h} - 1); 1 \leq i \leq N, 3 \leq l \leq M, 1 \leq h \leq l-2 \quad (7)$$

$$x_{i,k} \geq (y_{i,k-1} + y_{i,k+1} - 1)(b_{k+1} - b_k); 1 \leq i \leq N, 2 \leq k \leq M-1 \quad (8)$$

$$\sum_{i=1}^n \sum_{h=1}^k x_{i,h} \geq v y_{n+1,k}; 1 \leq k \leq M \quad (9)$$

$$-A(1 - y_{n+1,k}) + \sum_{i=1}^n \sum_{h=1}^k x_{i,h} \leq u; 1 \leq k \leq M \quad (10)$$

$$\sum_{h=1}^k y_{i,h} - \sum_{h=1}^k y_{i+1,h} \geq 0; n+1 \leq i \leq N-1, 1 \leq k \leq M \quad (11)$$

$$\sum_{j=1}^n \sum_{h=k}^l x_{j,h} \geq v(y_{i,k} + y_{i+1,l} - 1); n+1 \leq i \leq M-1, 1 \leq k \leq M-1, k < l < M; \quad (12)$$

$$-A(2 - y_{i,k} - y_{i+1,l}) + \sum_{j=1}^n \sum_{h=k}^l x_{j,h} \leq u; n+1 \leq i \leq M-1, 1 \leq k \leq M-1, k \leq l \leq M \quad (13)$$

$$-A(1 - y_{N,k}) + \sum_{i=1}^n \sum_{h=k}^M x_{i,h} \leq u; 1 \leq k \leq M \quad (14)$$

The objective function (2) is to minimize the total power cost of order processing, TEC. Constraint (3) indicates that a job may be processed in multiple electricity tariffs intervals, but the sum of the time in each interval should be equal to the processing time of the job. Constraint (4) means that the processing time of a job in a certain electricity tariff interval cannot exceed the processing time of the job; if a job is not processed in a certain electricity tariff interval, the processing time of the job in the electricity tariff interval should be 0. Constraint (5) imposes that the total processing time of all jobs in any electricity tariffs interval cannot exceed the length of the electricity tariffs interval, and ensures that the job processing does not overlap. Constraint (6) means that for any two adjacent electricity tariffs intervals k and $k+1$, at most one job can be processed in both electricity tariffs intervals at the same time, which restricts that only one job can be processed across the boundary between electricity tariffs intervals. Constraint (7) means that if a job is processed across multiple electricity tariffs ranges, these electricity tariffs ranges must be connected; constraint (8) means that only if the job processing time is greater than a certain electricity tariffs range, will the job cross the electricity tariffs range; at the same time; Constraints (7) and (8) ensure that the continuous processing of the job is not interrupted.

Constraint (9) and constraint (10) ensure the time interval requirement of the first maintenance, requiring the machine to process the job from time 0, and the total time for processing the job must not be less than v and not greater than u . Constraint (11) means that the i -th maintenance must be before the $i+1$ -th maintenance. Constraint (12) and constraint (13) ensure the time interval requirement for the i -th maintenance, the total time of processing jobs between intervals must not be less v than and not greater than u . Constraint (14) requires that the job processing time after the last maintenance work cannot be greater than u , otherwise this maintenance is not the last one.

3 Memetic algorithm

3.1 Algorithm design ideas and structure

Most of single scheduling problems under TOU mechanism are difficult to solve, which have been proved to be a typical NP-Hard problem by (Fang et al., 2016). Considering the complexity of our model integrating production and maintenance under TOU, this paper proposes an MA solution algorithm. The proposed MA uses a GA-based search framework. GA has the characteristics of good robustness and strong versatility, and is an effective algorithm for solving many scheduling problems (Wu, 2018).

However, the GA is easy to fall into a local optimum (R. Zhang & Chiong, 2016). This paper embeds two local search algorithms, TS and SA into the GA framework, and proposes an MA solution algorithm. The algorithm has the following three innovations:

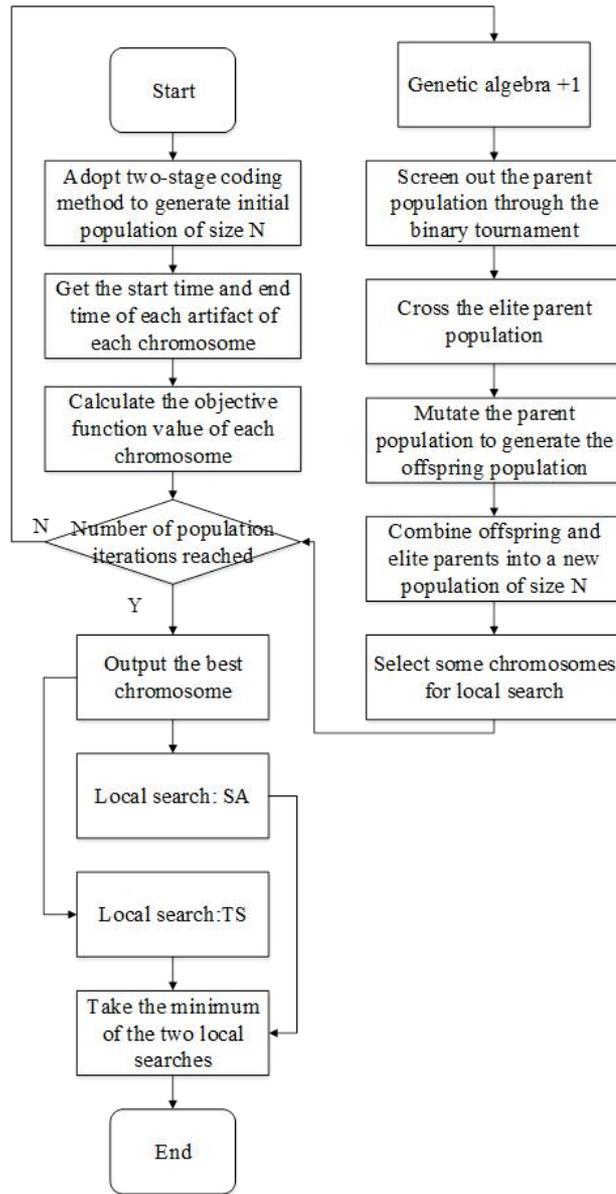


Figure 2: Flowchart of the MA.

(1) Parallel local search framework. On the basis of the results of the GA, the local optimal solution can jump out through the parallel local search enabled by TS and SA to reduce the error between the solution result and the optimal solution.

(2) Stacking adjustment strategy. Under the same job processing sequence, the difference in electricity prices is mainly caused by the different time intervals between the jobs. The local search algorithm is inserted in the algorithm solving process to improve the quality of chromosomes, and the time interval of adjacent jobs is dynamically adjusted. Compare the value of the objective function before and after adjustment to find a better chromosome.

(3) Penalty function-based search strategy. Due to the limitation of maintenance work, it is difficult to obtain qualified initial chromosomes when the problem is large. For this reason, the algorithm adds a penalty function in the decoding process to detect whether the total time length of the job between each two maintenance tasks meets the length of the maintenance interval. If it is not satisfied, the corresponding penalty value is added to the objective function value to reduce the probability of a chromosome being inherited.

The GA contains the time arrangement and sequence of the artifacts of the scheduling result information. In our algorithm, the sequence of all the artifacts and the start time of work need to be obtained through coding. If the sequence of job processing and the length of the interval between two adjacent jobs are determined, and the processing time of each job is combined, the starting time of each job in the scheduling time period can be obtained.

3.2 GA-based search framework

(Cui et al., 2019) used two-stage coding for job shop scheduling problems under TOU. In a scheduling problem with n processing job, the length of the encoding chromosome is $2n+1$, the 1 to n positions are the number of the randomly generated jobs, and the $n+1$ positions behind represents the time interval between two randomly generated jobs that are adjacent to each other, and the sum of all intervals should be equal to the total scheduling time minus the necessary processing time for all jobs, we called it is EM . The sequential combination of jobs processing time and interval can further obtain the start processing time and end time of each job, which is used to solve the electricity cost of each job in the sorting.

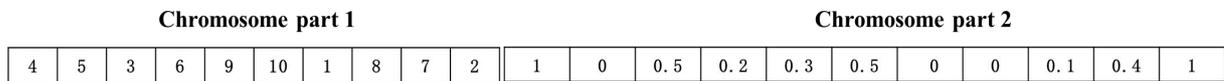


Figure 3: Coding diagram.

The algorithm uses binary tournament selection operator. First, we randomly selected two chromosomes A and B from the population, and chose the chromosome with a small objective function value to enter the elite parent population. After n times selection, the elite parent population of size n is obtained, and the offspring population is obtained through crossover of this population. Because the coding of the initial population adopts a two-stage format, the two preceding and following genes cannot cross each other. For the previous paragraph genes, we randomly generate a Boolean variable code of length n , and place the gene at the corresponding position of the A chromosome at the position where the corresponding code of the Boolean variable is 1, which is the number of a certain job. Then we search and delete the job codes that have been placed in the B chromosome. If there are i positions with the code 1, there are $n-i$ position that needs to be selected from the B chromosome and placed in the progeny population. Finally, we randomly generate a probability number P , and multiply all the second part genes on the A chromosome with P , all the second part genes on the B chromosome with $P-1$, and add the corresponding positions of the second part genes of the two chromosomes to obtain the offspring population. Get the interval between new artifacts in the offspring population. Repeat this method to $N/2$ times get the number of progeny population.

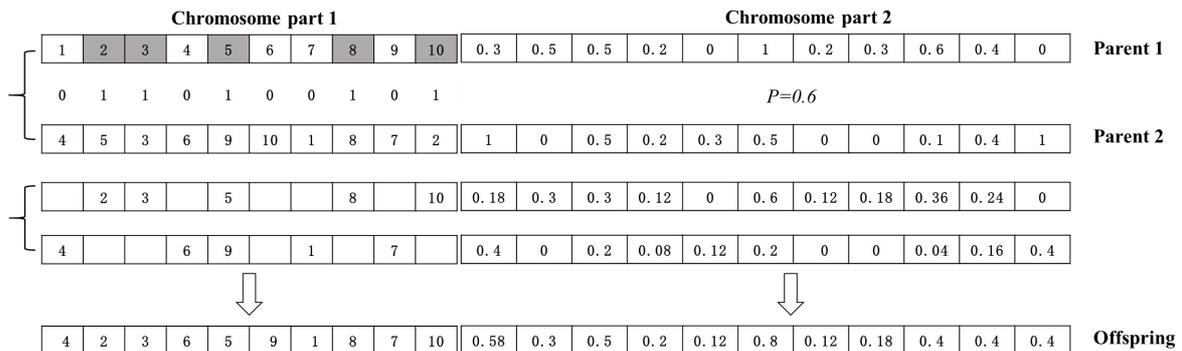


Figure 4: Crossover operator.

3.3 Stacking adjustment strategy

In the process of GA, the adjacent time interval of all chromosomes will be adjusted every fixed algebra, and the interval will be expanded as much as possible during the period of higher electricity tariffs. Processing and production will not be carried out in the interval of high electricity tariffs. The interval in the time period should be 0 as much as possible, and the job should be arranged in the time period with low electricity tariffs as much as possible. After many experiments, it was found that this method can quickly reduce the minimum energy cost of chromosomes. However, it is necessary to test every time interval of all chromosomes, which takes a long time, and the effect of frequent detection and adjustment is not obvious. We choose to adjust every 20 or 10 generations.

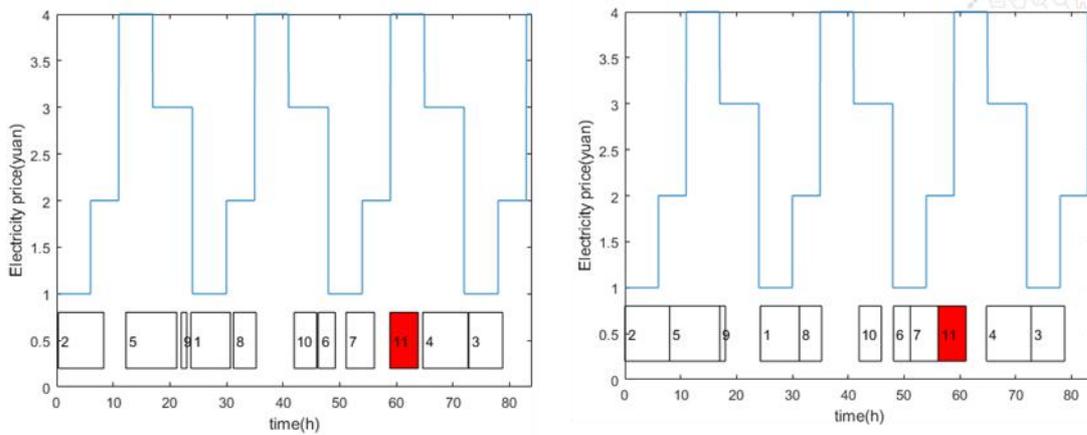


Figure 5: Stacking adjustment strategy.

3.4 Local search design

This paper designs a local search scheme to improve the quality of the GA. When the GA completed, the local search base on the optimal chromosome can quickly obtain a solution with a smaller error. The SA and TS will be used to realize local search.

The simulated annealing algorithm needs to set the initial temperature, randomly exchange genes at different positions on the target chromosome or mutate the processing time interval of the job to obtain a new chromosome, and determine the probability of accepting a poor solution according to the value of the objective function and the change of temperature. When the value of the objective function changes from $E(n)$ to $E(n+1)$, the probability formula for accepting the new solution is as follows.

$$P = \begin{cases} 1, & E(n+1) < E(n) \\ e^{-\frac{E(n+1) - E(n)}{T}}, & E(n+1) \geq E(n) \end{cases} \quad (15)$$

In the early stage, the initial temperature is higher, there is a high probability that a worse solution will be accepted, so that it can jump out of the local optimum. As the number of algorithm operations increases, the temperature gradually decreases, and the probability of

accepting poor solutions becomes smaller and smaller. The temperature change formula is as follows. We set the parameter λ to 0.99.

$$T(n+1) = \lambda T(n), n = 1, 2, 3, \dots \quad (16)$$

The crucial stage of TS is to set the tabu table. In each generation of mutation, the mutation method with the smallest objective function value is selected as the initial chromosome for the next operation, and this method is recorded in the tabu table. After that, it cannot be selected within the length of the tabu table. In this way, the search is prevented from falling into the local optimum. The length of the tabu table is related to the scale of the problem. If the tabu length is too short, the loop will not be able to break out of the local best point; if the tabu length is too long, all candidate solutions will be tabu, resulting in a long calculation time, which may make the calculation impossible to proceed. The formula for setting the length (L) of the tabu table in this article is as follows.

$$L = \sqrt{2A(2A+1)} \quad (16)$$

Finally, the smaller value obtained by the two methods is selected as the optimal solution.

4 Numerical study

4.1 Algorithm performance

In order to verify the effectiveness and accuracy of the design algorithm, we used CPLEX to solve the model and compare the optimal solution with the algorithm result. We set up the number of jobs from 10 to 100 kinds of scale problem calculation examples, the time period of TOU is [6,5,6,7] four time periods (unit: hour), and the electricity price of each time period is [1,2,4,3] (Unit: Yuan). According to the change data of the TOU in a period, the total scheduling time length is divided into several periods of periodic change.

The experimental platform is based on Lenovo Tianyi 510Pro, Intel i7-9700 3.0 GHz, 16GB RAM. The algorithm is written in MATLAB R2018b, and the CPLEX solver is called through the java program to calculate the optimal solution of the example. The optimal solution is compared with the algorithm solution result to discuss the advantages and disadvantages of the algorithm.

Table 1: This is an example of a table

n	v	u	EM	CPLEX time(s)	MA		SA	TS
					times(s)	error	error	error
10	12	48	24	0.09	1.82	2.32%	3.68%	2.76%
	12	24	24	0.17	2.14	3.66%	7.44%	3.94%
	12	24	72	2.93	2.83	2.36%	5.88%	2.80%
30	24	48	36	157.79	14.39	1.69%	2.10%	2.18%
	24	48	72	213.00	14.80	1.33%	2.13%	1.52%
	24	72	72	342.97	13.97	2.52%	3.61%	2.77%
50	24	48	24	\	63.20	\	\	\
	24	48	48	\	61.74	\	\	\

According to the experimental results, when the scale of the calculation example is small, CPLEX can quickly solve the optimal solution. The improved GA algorithm can obtain an approximate solution with an error within 6%, and the error can be further reduced to 1% to 3% after the SA and TS domain search algorithms. When the scale of calculation examples gradually increases, the speed advantage of the algorithm is reflected. CPLEX may not find suitable upper and lower bounds for a long time. The algorithm's solving speed will not undergo sudden changes, and as the scale of the calculation example increases, the genetic algebra will increase slowly.

Table 2: The impact of key parameters on algorithm performance

n	v	u	EM	CPLEX time(s)	MA		SA	TS
					times(s)	error	error	error
10	12	48	24	0.09	1.82	2.32%	3.68%	2.76%
	12	24	24	0.17	2.14	3.66%	7.44%	3.94%
	12	24	72	2.93	2.83	2.36%	5.88%	2.80%
30	24	48	36	157.79	14.39	1.69%	2.10%	2.18%
	24	48	72	213.00	14.80	1.33%	2.13%	1.52%
	24	72	72	342.97	13.97	2.52%	3.61%	2.77%
50	24	48	24	\	63.20	\	\	\
	24	48	48	\	61.74	\	\	\

According to the analysis of the experimental results, the size of the maintenance interval and the length of the idle time have a greater impact on the solution speed. In order to verify the size of the maintenance interval and the length of the idle time, the following sensitivity analysis test is set up, keeping other parameters unchanged under the same job size and parameters, changing a single variable, and analyzing the impact of this variable on the performance of the algorithm. The results are as follows.

4.2 Sensitivity analysis

4.2.1 The effect of due date on TEC

To analyze the impact of TEC in the production process, this paper comprehensively considers the due date, maintenance intervals setting and the length of the maintenance window and other factors, and sets up related sensitivity experiments.

In the first experiment, we analyze the impact of due date on TEC. The experiment data shows that when the jobs processing parameters are unchanged and the due date gradually increases, the TEC first reduces then keep consistent. After analysis, we learned that when the due date increases, the length of the lower electricity p periods also increases. Under the condition of ensuring maintenance constraints, more jobs can be arranged in the low electricity periods, TEC will become lower. When the due date further increases and all jobs can be arranged in the lowest electricity periods, the TEC drops to the minimum. TEC will not continue to drop. This will help managers to determine whether urgent orders are accepted taking into account total energy consumption.

Table 3: The effect of due date on TEC

$n = 20 \quad T = 60$		$n = 10 \quad T = 29$		$n = 10 \quad T = 29$	
d_{\max}	TEC	d_{\max}	TEC	d_{\max}	TEC
84	484	53	184	43	192
108	425	77	165	55	163
132	389	101	160	67	158
156	364	125	159	79	146
180	345	149	159	91	146
204	335	173	160	103	134
228	332	197	159	115	134
252	329	221	159	127	132
276	328	245	159	139	132
300	328	269	159	151	132

4.2.2 The effect of maintenance intervals on TEC

Then we analyze the impact of maintenance intervals on energy consumption. According to the previous analysis, we know that the maintenance interval setting will affect the number of maintenances required during the operation of the machine and the maintenance schedule range. While keeping other production factors unchanged, When the length of the maintenance interval is fixed, as the lower limit increases, the TEC will first increase and then decrease. When the lower limit is fixed, as the interval length becomes longer, the TEC gradually decreases. On the one hand when the length of the maintenance interval remains unchanged and the lower bound of the maintenance interval initially becomes larger, it limits the maintenance that cannot be scheduled during the period of high electricity prices in the early period, resulting the TEC increase. When the lower bound of the maintenance interval is further enlarged, the upper bound will increase accordingly, which will reduce the maintenance number, the TEC will decrease. On the other hand when the lower bound of the maintenance interval remains unchanged and the length of the maintenance interval increases, it means that the maintenance schedule is more flexible and the upper bound of the

maintenance interval increases. The maintenance frequency is reduced and maintenance can be scheduled in a larger scope. Before the due date, the maintenance can be arranged more independently, which reduces the TEC.

Table 4: The effect of maintenance intervals on TEC

$n = 30 \quad T = 84$				$n = 50 \quad T = 149$			
v	u	L	TEC	v	u	L	TEC
24	48	24	590	24	48	24	1290
36	60	24	591	36	60	24	1297
48	72	24	602	48	72	24	1314
60	84	24	602	60	84	24	1301
72	96	24	588	72	96	24	1321
36	84	48	591	36	84	48	1289
48	120	72	590	48	120	72	1293
60	156	96	590	60	156	96	1299
72	192	120	588	72	192	120	1292

The finding obtained by analyzing the above factors can provide managers with suggestions for better management and maintenance arrangements. It benefits reduction of energy costs and operational expenses.

5 Conclusion

This paper considers an energy-efficient scheduling problem under TOU incorporating integrated scheduling decisions of flexible periodic preventive maintenance of production equipment. With the goal of minimizing TEC, we establish a mixed integer program model taking into account due date of customer order.

We design an improved MA based on GA-based population search combined with TS and SA. By comparing the experimental results of the MA with GA, MA can get better results than GA. Besides, the MA obtains near-optimal solutions (averagely 2.1%) compared to optimal solutions, requiring much less computation time compared by CPLEX.

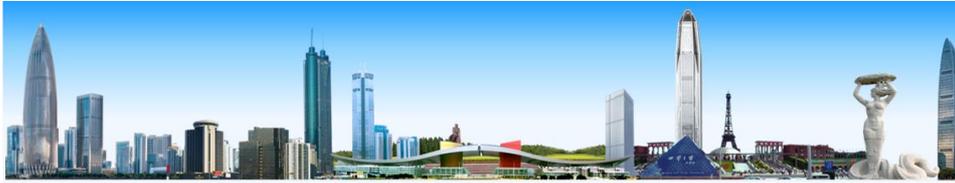
Finally, we analyze the key factors that affect energy consumption, including the due date, maintenance interval length and lower limit. We find that TEC first reduces then keep consistent as the due date increases. This will help managers to determine whether urgent orders are accepted taking into account total energy consumption. When the length of the maintenance interval is fixed, as the lower limit increases, the TEC will first increase and then decrease. When the lower limit is fixed, as the interval length becomes longer, the TEC gradually decreases.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (No. 51705250), Natural Science Foundation of Jiangsu Province (BK20170797), and China Postdoctoral Science Foundation (2019M661839). This research was also supported by the Nanjing University of Aeronautics and Astronautics Innovation Base (Laboratory) Open Fund (No. kfj20190908).

References

- Allaoui, H., Artiba, A., Goncalves, G., & Elmaghraby, S. E. (2008). Scheduling n jobs and preventive maintenance in a single machine subject to breakdowns to minimize the expected total earliness and tardiness costs. *IFAC Proceedings Volumes*, 41(2), 15843–15848. <https://doi.org/10.3182/20080706-5-KR-1001.02678>
- Che, A., Zeng, Y., & Lyu, K. (2016). An efficient greedy insertion heuristic for energy-conscious single machine scheduling problem under time-of-use electricity tariffs. *Journal of Cleaner Production*, 129, 565–577. <https://doi.org/10.1016/j.jclepro.2016.03.150>
- Cui, W., Sun, H., & Xia, B. (2019). Integrating production scheduling, maintenance planning and energy controlling for the sustainable manufacturing systems under TOU tariff. *Journal of the Operational Research Society*, 1–20. <https://doi.org/10.1080/01605682.2019.1630327>
- Fang, K., Uhan, N. A., Zhao, F., & Sutherland, J. W. (2013). Flow shop scheduling with peak power consumption constraints. *Annals of Operations Research*, 206(1), 115–145. <https://doi.org/10.1007/s10479-012-1294-z>
- Fang, K., Uhan, N. A., Zhao, F., & Sutherland, J. W. (2016). Scheduling on a single machine under time-of-use electricity tariffs. *Annals of Operations Research*, 238(1–2), 199–227. <https://doi.org/10.1007/s10479-015-2003-5>
- Luo, H., Du, B., Huang, G. Q., Chen, H., & Li, X. (2013). Hybrid flow shop scheduling considering machine electricity consumption cost. *International Journal of Production Economics*, 146(2), 423–439. <https://doi.org/10.1016/j.ijpe.2013.01.028>
- Moon, J.-Y., & Park, J. (2014). Smart production scheduling with time-dependent and machine-dependent electricity cost by considering distributed energy resources and energy storage. *International Journal of Production Research*, 52(13), 3922–3939. <https://doi.org/10.1080/00207543.2013.860251>
- Sharma, A., Zhao, F., & Sutherland, J. W. (2015). Econological scheduling of a manufacturing enterprise operating under a time-of-use electricity tariff. *Journal of Cleaner Production*, 108, 256–270. <https://doi.org/10.1016/j.jclepro.2015.06.002>
- Wu, X. (2018). A green scheduling algorithm for flexible job shop with energy-saving measures. *Journal of Cleaner Production*, 16.
- Zhang, H., Zhao, F., Fang, K., & Sutherland, J. W. (2014). Energy-conscious flow shop scheduling under time-of-use electricity tariffs. *CIRP Annals*, 63(1), 37–40. <https://doi.org/10.1016/j.cirp.2014.03.011>
- Zhang, Q., Manier, H., & Manier, M.-A. (2012). A genetic algorithm with tabu search procedure for flexible job shop scheduling with transportation constraints and bounded processing times. *Computers & Operations Research*, 39(7), 1713–1723. <https://doi.org/10.1016/j.cor.2011.10.007>
- Zhang, R., & Chiong, R. (2016). Solving the energy-efficient job shop scheduling problem: A multi-objective genetic algorithm with enhanced local search for minimizing the total weighted tardiness and total energy consumption. *Journal of Cleaner Production*, 112, 3361–3375. <https://doi.org/10.1016/j.jclepro.2015.09.097>



Physical Internet-enabled synchronized optimization for Milk-run transportation and Cross-docking warehouse in industrial park

Yuanxin Lin^{1,3,4}, Ting Qu^{2,3,4}, Kai Zhang^{1,3,4} and George Q Huang^{3,5}

1.School of Management, Jinan University, Guangzhou, PR China

2.School of Intelligent Systems Science and Engineering, Jinan University (Zhuhai Campus), Zhuhai, PR China

3.Institute of Physical Internet, Jinan University (Zhuhai Campus), Zhuhai, PR China.

4.Institute of the Belt and Road & Guangdong-Hong Kong-Macao Greater Bay Area, Jinan University, PR China

5.Dept. Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Hong Kong, China

Corresponding author: Ting Qu address email: quting@jnu.edu.cn

Abstract: For the complex system of the industrial park consisting of the three stages of manufacturing, transportation, and warehouse, the increasingly lean and intelligent manufacturing system can gradually meet the increasing customized demand. However, if the transportation and warehousing stages with a low degree of intelligence cannot effectively match the manufacturing process in real-time, it will cause not only an increase in overall operating costs but also a decline in customer service levels. This paper focuses on the finished product warehousing process in an industrial park under a highly dynamic production environment, and studies the synchronized decision-making problem of transportation fleet and finished product warehouse under the Physical Internet (PI) environment. A PI-based synchronization information framework is proposed to solve the challenge of state perception. To solve the challenge of decision level, Collaborative Optimization (CO) is used to systematically coordinate and optimize the "transportation-warehousing" units with independent decision-making and synchronized operations. Finally, simulation and comparative analysis of the proposed synchronization solutions were carried out based on the actual production data of the cooperative enterprises. The results showed that the proposed method can improve the utilization rate of system resources and reduce the operation cost.

Keywords: Physical Internet; synchronization; milk-run; cross-docking; collaborative optimization

1 Introduction

To meet the optimal response to customized demand, manufacturers are gradually adopting flexible production models, while wholesalers/retailers expect to achieve small-batch deliveries in a short time window. Having very little inventory is an important feature of customized production. Many works of literature believe that there is no finished product inventory under customized production mode (Carr et al. 2000 & He et al. 2002). In fact, customer orders include multiple varieties and small batches of products produced by different manufacturers, which requires a certain space to temporarily store and integrate products of the same order. However, due to the low and peak seasons of market demand, to avoid investment in capital, manpower, and resources, many manufacturers tend to look for third-party public logistics services instead of building their warehousing systems. The Supply Hub in Industrial Park (SHIP) is a public warehouse that effectively integrates transportation and storage resources. It can provide raw material/finished product logistics services for multiple manufacturers to achieve Just-in-Time (JIT) operations (Qiu et al. 2012). Gradually, transportation companies

have shifted their strategy from a large-volume, large-scale transportation mode to Milk-run (MR) to achieve small-volume, high-frequency transportation demand. Warehousing operations will also shift from long-period storage to short-period Cross-docking (CD), realizing the temporary storage of finished products and the integration of different products in the same order (Luo et al. 2019).

Under customized demand, the dynamics of orders, resources, and processes will cause the operation process to be disconnected, resulting in a waste of resources and delays in delivery. For this reason, SHIP has to consider: (1) How to realize batch dynamic transportation, which can improve transportation cost efficiency while ensuring the completeness of the products in the order; (2) How to achieve unitized dynamic storage, which can reduce the storage time of orders while increasing the utilization rate of the warehouse, and shorten the delivery lead time as much as possible; (3) More attention should be paid to the fact that if the transportation only pursues the scale of transportation, it is easy to increase the overall transit time of the product; if the warehouse only pursues the turnover efficiency, it is easy to cause waste of transportation resources. Therefore, how to adjust the two plans to keep the time parameters consistent, to achieve the lowest overall operating cost on the premise of meeting the product delivery time and customer demand, which became the research problem of this paper.

Production logistics synchronization (PLS) is to establish an association between two or three subsystems of independently distributed production, transportation, or warehouse models, so that they can achieve synergy in resource allocation, execution plan, control parameters, etc. (Qu et al. 2016). This paper takes the finished product warehousing process in the industrial park as the research object and defines the synchronization problem between MR transportation and CD warehouse as MR-CD Synchronization. At present, a variety of cutting-edge (e.g., Physical Internet) can be used to solve MR-CD Synchronization in industrial parks. The concept of Physical Internet (PI) is developed based on the Internet of things (IoT) and cloud computing technology (Kong et al. 2012). Although there is a lot of research in this area, there are still some limitations. These research gaps will be transformed into the following research questions: (1) At the perception level, construct a synchronization information framework to realize the real-time perception of the status of the manufacturing, transportation, and warehouse stages and feedback to the decision-making layer for decision-making; (2) At the decision-making level, synchronization control mechanism and method are established to coordinate the inconsistency between transportation and warehouse targets, and the control instructions are issued to the frontline to achieve closed-loop control.

The remaining of this paper is organized as follows. Section 2 analyzes the problem of the finished product warehousing process in the industrial park. Section 3 proposes a solution for synchronized decision-making. Section 4 carries on the case analysis. The last section summarizes the paper and identifies potential future works.

2 Analysis of synchronization problems

2.1 Operation process

The storage process of finished products in the industrial park mainly involves two types of enterprises, manufacturers, and SHIP. As shown in Figure 1, SHIP includes SHIP Fleet and SHIP Warehouse.

Business process: The manufacturer relies on the transportation and warehousing services provided by SHIP to realize the transfer and storage of finished products to reduce the capital investment in logistics and warehouse construction. SHIP integrates the demand of manufacturers to formulate cargo collection and warehousing plans reasonably allocates pickup

vehicles and storage locations and ensures that the delivery time of products meets customer demand.

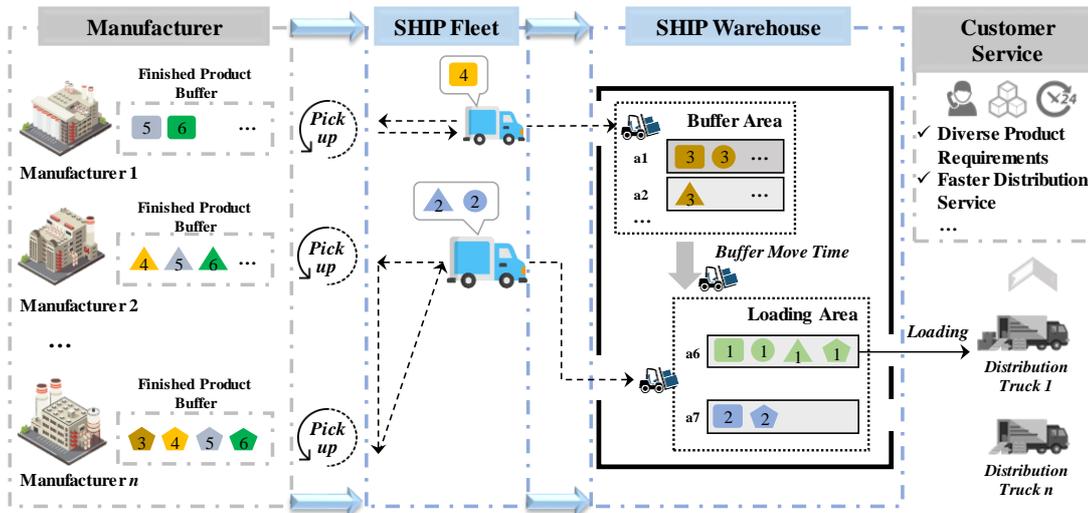


Figure 1: Finished product warehousing process in the industrial park

2.2 Problem analysis

2.2.1 Difficulties in local decision making

- For the transportation stage

What SHIP Fleet considers is the Milk-run transportation decision problem, which has the following characteristics: (1) Generally, the combination of product types in customer orders is complicated. Manufacturers make their production plans based on product demand, resource constraints, etc. Therefore, the product's off-line time will be different; (2) Manufacturers are located in different locations in the industrial park. For some large industrial parks, they may be further apart; (3) To meet the transportation demand of various batches, SHIP Fleet has different types of vehicles with different rated capacities.

Therefore, in the face of factors such as unsynchronized completion time, complex driving paths, and diverse transportation resources, how to plan the types and driving path of the pickup vehicle with the lowest transportation cost has become a difficult point in transportation decision-making.

- For the warehouse stage

What SHIP Warehouse needs to consider is the storage location assignment problem, which has the following characteristics: (1) Since the product stays in the SHIP warehouse for a relatively short time, to increase the circulation speed, the pallets are usually directly stored in the aisle on the ground. (2) SHIP warehouse is divided into two areas: the buffer area and the loading area. The aisle in the loading area is directly connected to the loading platform, and the aisle can be easily loaded onto the distribution trucks. If there is no available aisle in the loading area, the pallets have to be stored in the buffer area first.

Therefore, in the face of the limited storage capacity of the warehouse (especially in the peak sales season), how to plan the warehousing time of orders and allocate suitable aisle with the lowest storage cost has become a difficult point for warehouse decision-making.

2.2.2 Difficulties in synchronized decision-making

The decision-making results of SHIP Fleet and SHIP Warehouse are independent and mutually influencing. Under dynamic interference, there are the following operational problems:

- The mismatch between planning and execution: The tasks in each stage of the industrial park need to be executed according to the pre-made operation plan. However, due to the large number of random dynamics brought by customization demand, it is easy for the system to fail to complete the execution tasks according to the static decision results made under the ideal static state, or even deviate far.
- Poor synchronization of the execution process: (1) The backlog of finished products at the off-line point will cause the increase of factory operation costs;(2) Different decisions such as vehicle type, departure time, and the driving route will result in different SHIP Fleet transportation costs and finished product storage time;(3) The difference in the decision-making of the finished product's warehousing time and storage location will cause the SHIP warehouse storage cost and the finished product's storage time to be different. Because different companies (even different departments within the company) usually use different management information systems (e.g., ERP, TMS, WMS) for decision-making, performance accounting, etc. This is likely to cause the decision-making process of each stage to be a serial static local optimization process, lacking decision-making for global collaborative optimization with other stages.

2.3 Challenge analysis

As described above, the warehousing process of finished products in the industrial park involves multi-stage participation and multi-parameter interaction. In the dynamic customization production environment, the dynamics generated by any stage may cause inconsistent rhythm in the entire operation process. PLS requires that each decision-making stage be able to carry out autonomous and relevant dynamic coordination based on real-time information under dynamic interference, to eliminate or reduce the impact of random interference on the production system. Therefore, the following two challenges need to be solved.

- Condition monitoring: Under the customized production mode, the finished product warehousing process will inevitably be affected by various dynamic factors (e.g., order changes, resource failure, etc.). Therefore, an information acquisition method is needed to realize the real-time and accurate perception of information in a dynamic operating environment, which is an important foundation for implementing synchronized decision-making applications in the decision-making layer.
- Process control: Traditional experience-based management mechanisms and integrated decision-making methods are difficult to apply to highly dynamic production environments with complex resource relationships. Therefore, to maintain the global optimal operation of the system, it is necessary to introduce appropriate control mechanisms and coordinate optimization methods for guidance, and conduct real-time collaborative control according to the dynamics that occur in each correlation stage.

3 PI enabled MR-CD synchronization solutions

3.1 MR-CD synchronization information framework

The synchronization information framework of this paper is based on the AUTOM information framework solution proposed by Huang et al. (2012). The AUTOM framework is an extensible Manufacturing Internet of Things (MIoT) information framework that complies with the ISA-95 international standard. It provides a theoretical basis and a feasible way for the establishment of the IoT environment in industrial parks.

As shown in Figure 2, the information framework is composed of a smart object layer, gateway layer, service layer, and application layer. First of all, the smart object layer serves the frontline operation of the industrial park. Secondly, the gateway layer and the service layer are the basic equipment for realizing information transmission in the industrial park. Finally, the application layer provides decision support for manufacturers, SHIP, and other enterprises in the industrial park.

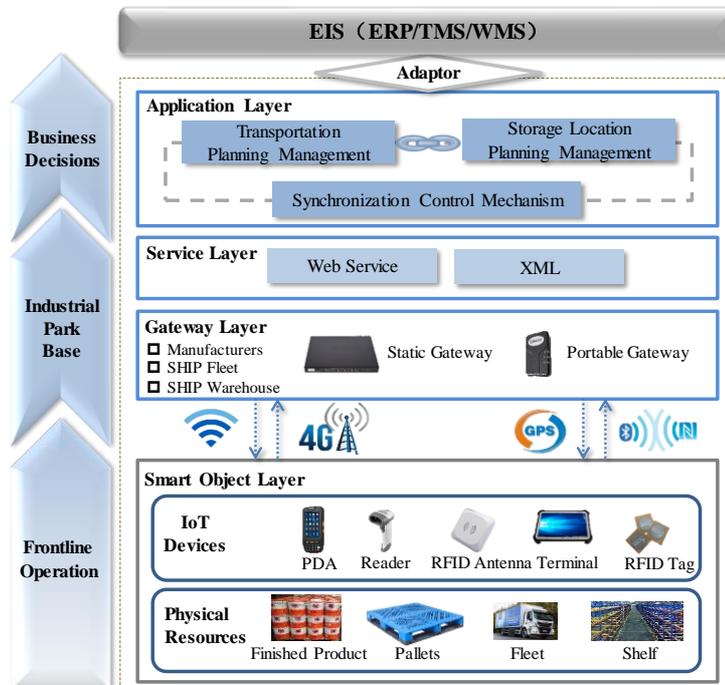


Figure 2: MR-CD synchronization information framework

3.2 MR-CD synchronization control mechanism

Based on the synchronization concept proposed by our research team (Qu et al. 2016; Zhang et al. 2020), considering the various dynamic impact ranges in the finished product warehousing process of industrial parks. This section proposes the MR-CD synchronization control mechanism from the qualitative perspective. The synchronization stage of this mechanism is divided into the plan-making phase and plan-correction phase, as shown in Figure 3.

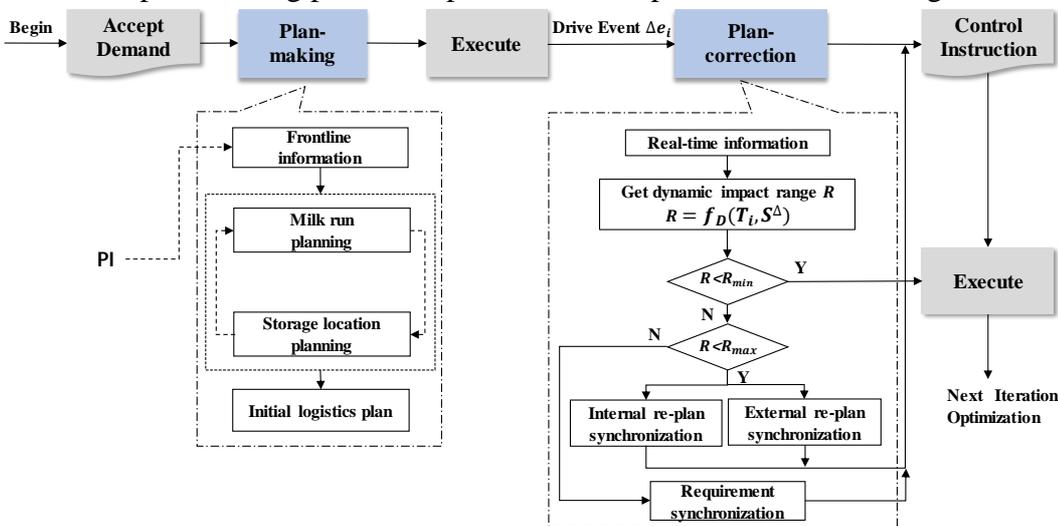


Figure 3: The synchronization control mechanism

- Plan-making phase

After receiving the manufacturer's delivery demand, basic data such as vehicles, order delivery dates, and cargo lane capacity are obtained through the smart object layer. At the same time, SHIP makes collaborative decisions on the transportation plan and the warehousing plan until the initial logistics plan with the lowest overall operation cost is produced.

- Plan-correction phase

During the operation of the production system, the occurrence of various dynamic events may lead to deviations between actual execution and planned data. First, the synchronization control mechanism judges the dynamics of the execution layer. Then, the corresponding synchronization mode is triggered according to the dynamic range of influence. Finally, after formulating a revised plan, it is issued to the executive layer in the form of control instructions.

3.3 Collaborative optimization

MR-CD synchronization is an optimization problem involving the two disciplines of transportation and warehouse. It has the characteristics of distributed and coupled. This is mainly reflected in the fact that all decision-making bodies are at the same level and interdependent. Collaborative Optimization (CO) is one of the multidisciplinary optimization methods to solve large-scale complex coupling problems, which has been introduced into the optimization of production systems by many scholars (Qu et al. 2017; Lin et al. 2020). The CO algorithm has a typical two-level optimization structure, so it is suitable for solving the MR-CD synchronization problem in this paper.

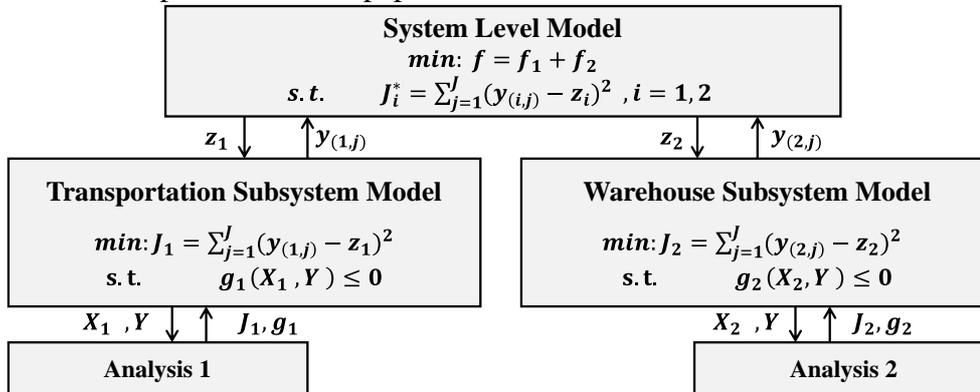


Figure 4: CO framework

Figure 4 shows the collaborative optimization framework of CO. The top layer of the framework is the optimization of logistics system-level disciplines, and the bottom layer is the optimization of two subsystems, transportation, and warehouse. The entire synchronized decision-making problem is decomposed into two-level nonlinear programming problems: (1) The system level allocates design variables $z_i = \{X_i, Y\}$ to the transportation and warehouse subsystems respectively; (2) Under the condition of satisfying their respective constraints, the transportation and warehouse independently solve the optimal solution $y_{(i,j)}$ with the smallest possible deviation from z ; (3) The system level coordinates the coupling variable Y between transportation and warehouse through consistency constraints; (4) After multiple iterations of optimization, a solution that satisfies the consistency constraints and has the lowest logistics cost is obtained.

4 Case study

Based on the actual production data of a large chemical group in China that our research team cooperates with, this section verifies the synchronization solution proposed in this paper through data simulation.

4.1 Business process and basic data

The group has a total of four factories in its industrial park, producing wood paint, interior wall paint, exterior wall paint, and latex paint. The finished product warehousing process is similar to Figure 1. After the group customer service receives the customer's order, it will be assigned to the corresponding factory for production in the form of a production order according to the product type. Table 1 shows the production data of the group on a certain day after the desensitization treatment. In this paper, 10 am is set as 0 times. The completion time and the latest pickup time constitute the pickup service time window. Table 2 is the basic information of SHIP Fleet vehicles, and Table 3 is the basic information of SHIP Warehouse aisle.

Table 1: Customer order information

No	Customer Order	Manufacturer	Completion Quantity	Completion Time	The Latest Pickup Time	Delivery Time
1	1	1	6	168	190	260
2	1	2	8	129	161	260
3	1	3	10	210	239	260
4	1	4	10	114	143	260
5	2	1	5	147	179	350
6	2	2	6	144	177	350
7	2	3	12	300	333	350
8	2	4	8	156	181	350
9	3	1	8	345	379	470
10	3	2	10	292	309	470
11	3	3	6	333	366	470
12	3	4	13	243	277	470
13	4	1	6	96	134	200
14	4	2	8	72	102	200
15	4	3	8	168	196	200
16	4	4	10	96	127	200
17	5	1	8	249	284	380
18	5	2	10	244	269	380
19	5	3	5	315	346	380
20	5	4	15	218	249	380

Table 2: Basic data of vehicle

Vehicle Type	Loading Weight (pallets)	Number	Running Speed (km/h)
1	30	6	15
2	40	4	15

Table 3: Basic data of aisle

Aisle Type	Aisle Number	Area	Capacity(pallets)	Buffer Move Time(min)
1	a1-a5	Buffer area	20	30
2	a6-a7	Loading area	60	0

4.2 Mathematical models

According to the MR-CD Synchronization studied in this paper, this section proposes a CO-based synchronization optimization mathematical model from a quantitative perspective. To simplify the problem without loss of generality, the problem assumptions and parameter descriptions are shown in Tables 4 and 5. Among them, the units of the order quantity and the vehicle capacity mentioned in Table 4 and Table 5 are pallets.

Table 4: Problem assumption

No	Assumption	No	Assumption
1	There are several manufacturers in the industrial park, and each manufacturer is responsible for producing one type of product	2	A customer order contains multiple types of products, and different products are produced by the corresponding manufacturer
3	One category product in one order is a production order	4	The capacity of the finished product buffer is sufficient
5	Vehicle overload is not allowed	6	Movement time is fixed
7	The vehicle departs from SHIP and returns to SHIP after pickup	8	Ignore unloading and loading time of warehouse
9	Movement from the buffer area to the loading area is one-way	10	Production orders are not allowed to split for transportation
11	The number of pallets stored in the aisle cannot exceed its capacity	12	All products in the order will be delivered after entering the warehouse

Table 5: Parameter description

Notation	Description	Notation	Description
i	Manufacturer Number, $i = 1, 2, \dots, n$	C_{buf}^i	Penalty cost of delayed arrival of vehicle v
m	Customer order number, $m = 1, 2, \dots, M$	C_{fix}^v	Fixed cost of vehicle v
v	Vehicle number, $v = 1, 2, \dots, V$	C_{var}^v	Unit transportation cost of vehicle v
a	Aisle number, $a = 1, 2, \dots, N$	C_{fix}^a	Fixed storage cost of aisle a
Q_m	The total quantity of order m	C_{var}^a	Unit storage cost of aisle a
p_m^i	Quantity of production order	C_{del}	Delayed delivery time of order m
w_v	Loading weight of vehicle v	$t_{i,m}^{in}$	warehouse entry time of production order
T_0^v	The departure time of vehicle v	t_m^{in}	warehouse entry time of customer order m
t_{ij}	Travel time of vehicle v from i to j	T_{bm}	Buffer move time
at_i^v	The time when vehicle v arrives at i	dt_m^{req}	Delivery time of customer order m
$[et_i^k, lt_i^k]$	The k th time window of i	t_m^{out}	Allowed delivery time of customer order m
$wt_{i,m}^v$	Waiting time of vehicle v at i 1, if the production order is transported by vehicle v , and 0 otherwise	$[B_0, E_0]$	Operation time of ship
$g_{i,m}^v$		x_{ij}^v	1, if the vehicle runs from i to j , and 0 otherwise
s_i	Service time required by i	$\delta_{k,m}$	1, if order m is moved to loading area after k and 0 otherwise
q_i^v	Pick-up weight of i	η_m	1, if buffer area is the place order m located before it is transshipped to the loading area, and 0 otherwise
at_{war}^v	The time when vehicle v arrives at SHIP	$\theta_{m,a}$	1, if order m is located on aisle a , and 0 otherwise
V_a	The capacity of aisle a	$\mu_{m,i,a}$	1, if production order is located on aisle a , and 0 otherwise
C_{wait}^v	The waiting time cost of vehicle v	$u_{t,a}$	1, aisle a is used at time t , and 0 otherwise

4.2.1 System level model

The objective function and consistency constraints of the logistics system level are as follows:

$$\min f = f_t^2 + f_w^2 \quad (1)$$

$$J_t^* = (y_{(1,1)} - z_1)^2 \leq \varepsilon \quad (2)$$

$$J_w^* = (y_{(2,1)} - z_2)^2 \leq \varepsilon \quad (3)$$

Equation (1) represents the system level objective optimization function of the industrial park; Equations (2)-(3) represent consistency constraints, Where at_{war}^v is the coupling variable Y of the transportation subsystem model and the warehouse subsystem model.

4.2.2 Transportation Subsystem Model

$$\begin{aligned} \min f_t = & \sum_{v=1}^V \sum_{j=1}^n C_{fix}^v \times x_{0j}^v + \sum_{v=1}^V \sum_{i=1}^n \sum_{j=1}^n C_{var}^v \times t_{ij} \times x_{ij}^v + \\ & C_{wait}^v \sum_{v=1}^V \sum_{i=1}^n \max(et_i^k - at_i^v, 0) + C_{buf}^i \sum_{v=1}^V \sum_{i=1}^n \max(at_i^v - lt_i^k, 0) \end{aligned} \quad (4)$$

Equation (4) represents the transportation cost returned by the optimization of the transportation subsystem. The first term is the fixed cost of the vehicle, the second term is the running cost of the vehicle, and the third and fourth terms are the penalty cost of the time window.

$$\sum_{i=1}^n (q_i^v \sum_{j=1}^n x_{ij}^v) \leq w_v, v \in \{1, 2, \dots, V\} \quad (5)$$

$$\sum_{j=1}^n x_{0j}^v = \sum_{i=1}^n x_{i0}^v \leq 1, v \in \{1, 2, \dots, V\} \quad (6)$$

$$\sum_{v=1}^V \sum_{j=1}^n x_{ij}^v \geq 1, i \in \{1, 2, \dots, n\} \quad (7)$$

$$\sum_{m=1}^M p_m^i = \sum_{v=1}^V q_i^v, i \in \{1, 2, \dots, n\} \quad (8)$$

$$\sum_{v=1}^V g_{i,m}^v = 1, i \in \{1, 2, \dots, n\}, m \in \{1, 2, \dots, M\} \quad (9)$$

$$at_i^v + wt_{i,m}^v + s_i + t_{ij} = at_j^v \quad (10)$$

$$T_0^v + \sum_{i=0}^n \sum_{j=0}^n x_{ij}^v \times (wt_{i,m}^v + s_i + t_{ij}) = at_{war}^v \leq E_0, v \in \{1, 2, \dots, V\} \quad (11)$$

Equation (5) represents the overload constraint; Equation (6) represents that the vehicle departs from the SHIP and finally returns to SHIP; Equations (7)-(8) represents that the offline point allows multiple visits and the manufacturer's demand are met; Equation (9) represents that the production order is not allowed to be split and transported; Equation (10) calculates the time when the vehicle v arrives at the manufacturer j ; Equation (11) represents SHIP operating time constraints.

4.2.3 Warehouse Subsystem Model

$$\begin{aligned} \min f_w = & C_{fix}^a \sum_{a=1}^N u_{t,a} + \sum_{i=1}^n \sum_{m=1}^M (t_m^{out} - t_{i,m}^{in}) \times p_m^i \times C_{var}^a + \\ & C_{del} \sum_{m=1}^M \max(t_m^{out} - dt_m^{req}, 0) \end{aligned} \quad (12)$$

Equation (12) represents the storage cost returned by the optimization of the warehouse subsystem, the first and second terms are the fixed and variable costs of using the aisle, and the third term is the penalty cost of delayed delivery of customer orders.

$$\sum_{m=1}^M \mu_{m,i,a} \leq 2, a \in \{1, 2, \dots, N\} \quad (13)$$

$$\sum_{i=1}^n \sum_{a=1}^N \mu_{m,i,a} \leq 1, m \in \{1, 2, \dots, M\} \quad (14)$$

$$\sum_{m=1}^M \theta_{m,a} \leq 1, a \in \{1, 2, \dots, N\} \quad (15)$$

$$\sum_{a=1}^N \theta_{m,a} \leq 1, m \in \{1,2, \dots, M\} \quad (16)$$

$$\sum_{m=1}^M \sum_{i=1}^n \mu_{m,i,a} \times p_m^i \leq V_a, a \in \{1,2, \dots, N\} \quad (17)$$

$$\sum_{m=1}^M \theta_{m,a} \times Q_m \leq V_a, a \in \{1,2, \dots, N\} \quad (18)$$

$$t_{i,m}^{in} = at_{war}^v \sum_{v=1}^V g_{i,m}^v \quad (19)$$

$$t_m^{in} = \max(t_{i,m}^{in} \mid i = 1,2, \dots, n) \quad (20)$$

$$\sum_{m=1}^M \theta_{m,a} \times \mu_{m,i,a} = 0 \quad (21)$$

$$t_m^{out} = t_m^{in}, \text{ if } \eta_m = 0 \quad (22)$$

$$t_m^{out} = \sum_{k=1}^M \delta_{k,m} (t_m^{in} + T_{bm}), \text{ if } \eta_m = 1 \quad (23)$$

Equations (13)-(16) express the constraints of order storage in the cargo lane; Equations (17)-(18) express the capacity constraints of cargo lanes; Equations (19)-(20) calculate the arrival time of production orders and customer orders respectively; Equation (21) means that the products of a customer order can only be stored in the same area at the same time; Equations (22)-(23) respectively calculate the time for customer orders to meet the delivery demand.

4.3 Result analysis

Matlab R2016b software was used to program and simulate the above examples. The transportation and warehouse subsystem models are solved by genetic algorithm (GA). The GA parameters are set as follows: population size is 100, the iteration number is 500, crossover probability $P_c=0.6$, mutation probability $P_m=0.06$. At the same time, the CO method is used to coordinate the decision-making results of transportation and warehouse.

4.3.1 Plan-making phase

Before the start of production execution, driven by customer demand, static data on the frontline is obtained through the smart object layer. The application layer coordinates and optimizes the transportation decision and warehouse decision under the distributed coordination framework of the CO algorithm, to obtain the initial logistics plan with the minimum total system operation cost.

Table 6 shows the initial scheduling results of the SHIP Fleet. Among them, 0 represents SHIP, 1 in 1(2) represents the manufacturer number, and 2 represents the customer order number produced. Figure 5 shows the time when the pickup vehicle arrives at each manufacturer. It can be seen that the service can be reached within the time window of each manufacturer. Besides, Figure 6 is a Gantt chart for aisle distribution, which shows that customer orders can be shipped on time.

Table 6: The result of the initial route planning

No	Route	Vehicle Type	Loading Rate	Warehouse Entry Time
1	0-2(2)-1(2)-4(2)-3(4)-0	1	90%	172
2	0-1(1)-3(1)-4(5)-0	2	77.5%	222
3	0-2(5)-4(3)-3(3)-1(3)-0	2	92.5%	350
4	0-1(4)-4(1)-2(1)-0	1	80%	133
5	0-1(5)-3(2)-2(3)-3(5)-0	2	87.5%	320
6	0-2(4)-4(4)-0	1	60%	100

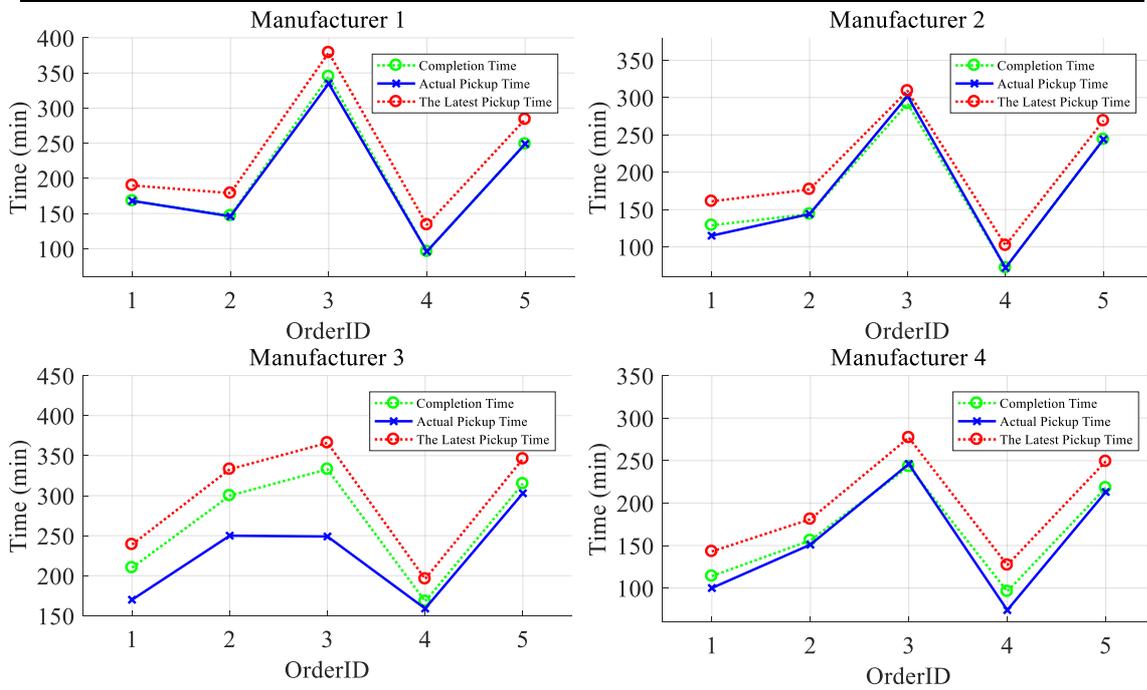


Figure 5: Initial planning of vehicle arrival time

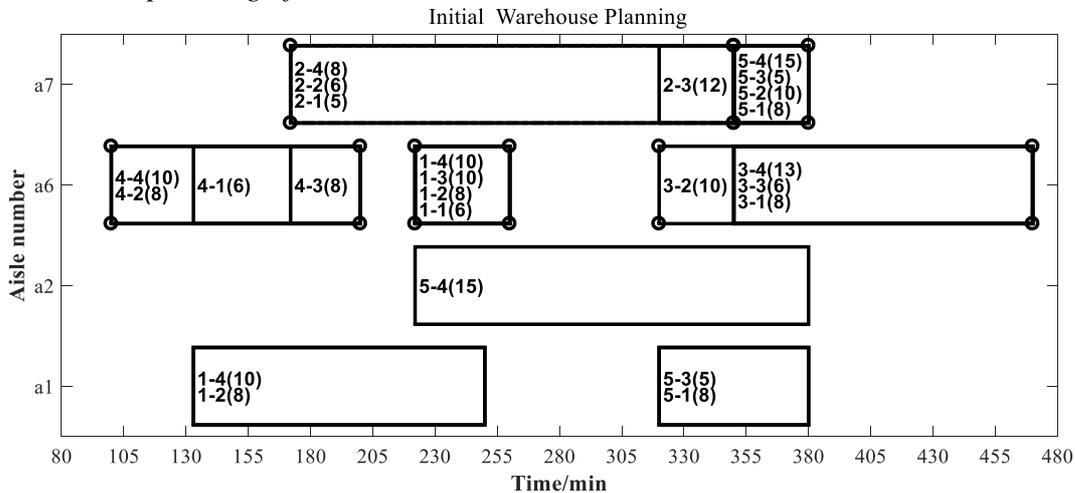


Figure 6: Gantt chart of initial warehouse planning

4.3.2 Plan-correction phase

This paper assumes that the chemical group received an urgent small-batch delivery order in the morning. The negotiated delivery time is 290, and the delivery number of each manufacturer is 4, 5, 8, 6, and the service time window is [180,207], [134,169], [228,266] and [68,98], respectively. This section compares the scheduling results with and without synchronized decision-making.

- Synchronization results

The addition of delivery orders will have an impact on the operation of the industrial park at all stages. At this time, the dynamics will trigger the MR-CD synchronization control mechanism, which is re-optimized through the CO algorithm. Table 7 shows the revised scheduling results of the SHIP Fleet. It can be seen from Figure 7 that the pick-up vehicles can still provide services to the manufacturers in time at the maximum loading rate. Besides, as shown in Figure 8, in the case of the least use of the aisle, both new and original orders can be delivered in time.

Table 7: The result of the corrected route planning

No	Route	Vehicle Type	Loading Rate	Warehouse Entry Time
1	0-4(2)-1(1)-3(4)-0	1	73.3%	174
2	0-2(5)-4(3)-3(2)-0	2	87.5%	304
3	0-4(6)-2(4)-4(4)-0	1	80%	100
4	0-2(1)-1(4)-2(6)-4(1)-2(2)-1(2)-0	2	100%	152
5	0-1(6)-3(1)-4(5)-3(6)-0	2	92.5%	232
6	0-1(5)-2(3)-3(3)-3(5)-1(3)-0	2	92.5%	350

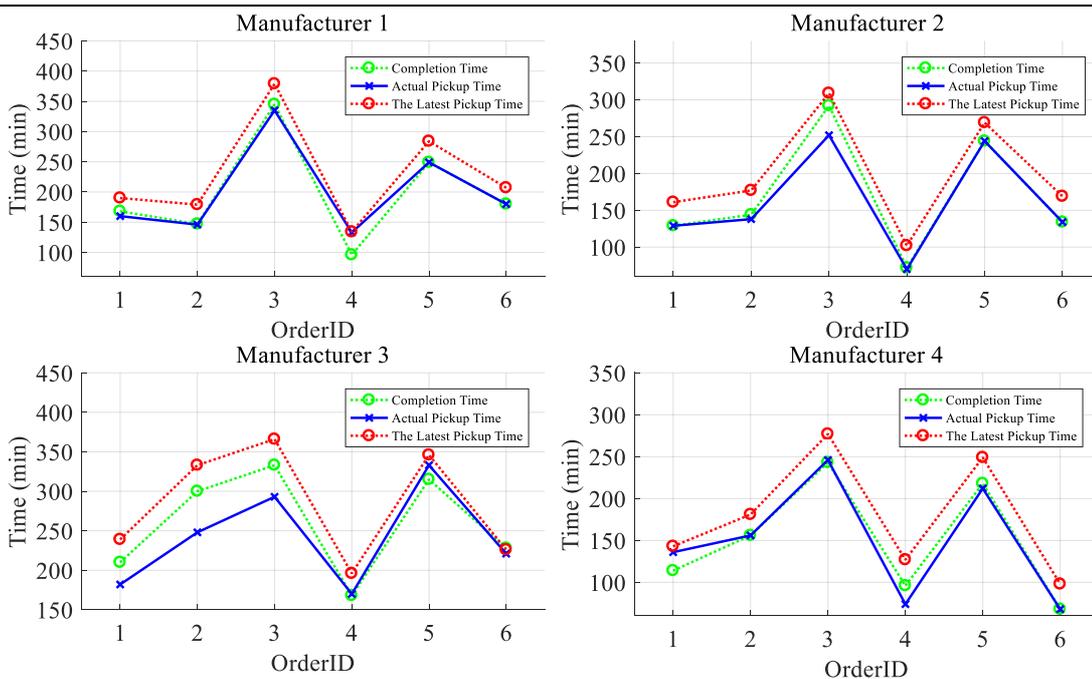


Figure 7: Corrected planning of vehicle arrival time

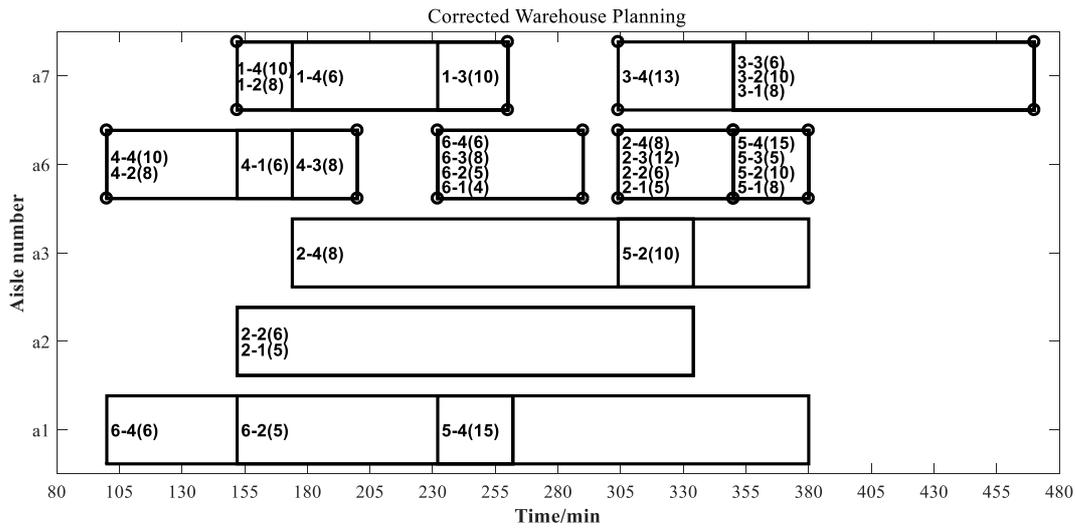


Figure 8: Gantt chart of corrected warehouse planning

- Comparative analysis of non-synchronization and synchronization results

Non-synchronization means that the newly added delivery orders are processed separately without changing the initial plan. Based on the initial logistics plan of subsection 4.3.1, the non-synchronization results are shown in Table 8 below.

Table 8: The comparison of results

Compared Items	Non-synchronization	Synchronization
Number of Vehicle Type 1/Type 2	4/3	2/4
Average Loading Rate	80.6%	87.6%
Number of Aisle Type 1/Type 2	3/2	3/2
Order On-time Delivery Rate	100%	100%

From the comparison in Table 7, it can be seen that the decision results of non-synchronization and synchronization can ensure that orders are delivered in time according to customer demand. Both SHIP Warehouse use the same number of the aisle. However, in the case of synchronized decision-making, SHIP Fleet can reasonably arrange to pick up vehicles at the maximum loading rate according to the results of the production offline, further reducing the overall operation cost. Therefore, the proposed solution can provide theoretical guidance for the process management of finished products in the actual industrial park.

5 Conclusion

This paper studies the MR-CD Synchronization problem that is crucial to improving the customer responsiveness of the industrial park system, and proposes PI enabled MR-CD synchronization solutions. To ensure that the system can still meet the delivery time required by customers after facing dynamic interference while ensuring the lowest overall operation cost. This synchronization solution can obtain real-time and accurate information based on the perception layer, to synchronized formulate the overall optimal logistics plan. Finally, the effectiveness of the program is verified through data simulation, and it can provide a feasible implementation framework and method for finished product logistics management in industrial parks.

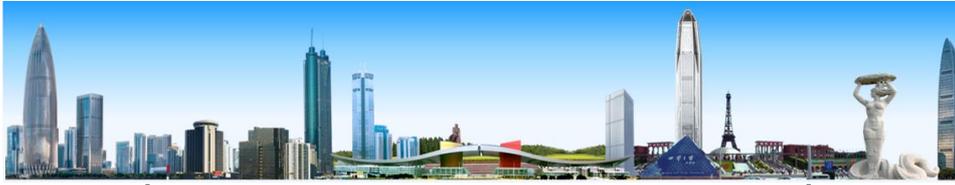
For future work, we will further study the overall perception of the real-time status of the system in combination with a complex dynamic production environment. Explore the synchronization control mechanism with production as the core.

Acknowledgement

This paper is financially supported from the National Natural Science Foundation of China (51875251), Guangdong Special Support Talent Program – Innovation and Entrepreneurship Leading Team (2019BT02S593), 2018 Guangzhou Leading Innovation Team Program (China)(201909010006), Blue Fire Project (Huizhou) Industry-University-Research Joint Innovation Fund of the Ministry of Education (China) (CXZJHZ201722), and the Fundamental Research Funds for the Central Universities (11618401). We would also like to thank Huizhou Jinze Industrial Development Co., Ltd. for their financial support to this project and the opportunity of system testing and implementing in their factories.

References

- Carr S., Duenyas I (2000): *Optimal admission control and sequencing in a make-to-stock/make-to-order production system*. Operations Research, v48, no5, 709-720.
- He, Q. M., Jewkes, E. M., Buzacott, J (2002): *Optimal and near-optimal inventory control policies for a make-to-order inventory-production system*. EUROPEAN JOURNAL OF OPERATIONAL RESEARCH, v141, no1, 113-132.
- Qiu X., Huang GQ (2013): *Supply Hub in Industrial Park (SHIP): The value of freight consolidation*. Computers and Industrial Engineering, v65, no1, 16-27.
- Luo, Hao., Yang, Xuan., Wang, Kai (2019) : *Synchronizedd scheduling of make to order plant and cross-docking warehouse*. Computers & Industrial Engineering, v138, 106108.
- Kong, XTR., Li, M., Yu, Y., Zhao, Z., Huang, GQ (2017). Physical internet-enabled e-commerce logistics park platform. 13th IEEE Conference on Automation Science and Engineering (CASE), Xian, China, August 2017.
- Huang, GQ., Zhang, YF., Jiang, PY (2008): *RFID-based wireless manufacturing for real-time management of job shop WIP inventories*. International Journal of Advanced Manufacturing Technology, v36, no7-8, 752-764.
- Qu, T., Lei, S., Wang, Z., Nie, D., Chen, X., Huang, George (2016) : *IoT-based real-time production logistics synchronization system under smart cloud manufacturing*. International Journal of Advanced Manufacturing Technology, v84, no1-4, 147-164.
- Zhang, Kai., Qu, Ting., Zhou, Dajian., Jiang, Hongfei., Lin, Yuanxin., Li, Peize., Guo, Hongfei., Liu, Yang., Li, Congdong., Huang, George (2020) : *Digital twin-based opti-state control method for a synchronizedd production operation system*. Robotics & Computer-Integrated Manufacturing, v63, 101892.
- Qu, T., Pan, YH., Liu, X., Kang, K., Li, CD., Thurer, M., Huang, GQ (2017) : *Internet of Things-based real-time production logistics synchronization mechanism and method toward customer order dynamics*. Transactions of the Institute of Measurement and Control, v39, no4, 429-445.
- Lin, YX., Qu, T., Zhang, K., Huang, GQ (2020) : *Cloud-based production logistics synchronization service infrastructure for customized production*. IET Collaborative Intelligent Manufacturing, v2, no3, 115-122.



Place your logo here

IPIC 2020 | 7th International Physical Internet Conference | Shenzhen

Multi-agent reinforcement learning-based dynamic task assignment for vehicles in urban transportation physical internet

Wei Qin, Yanning Sun, Zilong Zhuang, Zhiyao Lu and Yaoming Zhou

School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Corresponding author: wqin@sjtu.edu.cn

Abstract: *The transportation task assignment for vehicles plays an important role in city logistics of the physical internet, which is the key to cost reduction and efficiency improvement. The development of information technology and the emergence of “sharing economy” create a more convenient logistics mode, but also bring a greater challenge to efficient operation of urban transportation physical internet. On the one hand, considering the complex and dynamic environment of urban transportation, an efficient method for assigning transportation tasks to idle vehicles is desired. On the other hand, to meet the users’ expectations on immediate response of vehicle, the task assignment problem with dynamic arrival remains to be resolved. In this paper, we proposed a dynamic task assignment method for vehicles in urban transportation physical internet based on the multi-agent reinforcement learning. The transportation task assignment problem is transformed into a stochastic game process from vehicles’ perspective, and then an extended actor-critic algorithm is proposed to obtain the optimal strategy. Based on the proposed method, vehicles can independently make decisions in real time, thus eliminating a lot of communication cost. Compared with the methods based on FCFS (first come first service) rule and classic contract net, the results show that the proposed method can obtain higher acceptance rate and average return in the service cycle.*

Keywords: *urban transportation physical internet, transportation task assignment, multi-agent reinforcement learning, actor-critic algorithm.*

1 Introduction

With the increasingly fierce market competition and the advancement of information technology, the existing city logistics modes are developing towards an energy-saving, efficient and shareable manner. In particular, the novel mode combining city logistics with physical internet, so-called hyperconnected city logistics (Ballot et al., 2014; Kubek and Więcek, 2019;), makes traffic management system to operate more effectively by big data analysis and machine intelligence algorithms (Zhong et al., 2017; Kaffash et al., 2020). In this kind of traffic management system, assigning transportation tasks to vehicles is one of the most important services. Rapidity and rationality are the guarantee for the satisfaction of both users and drivers. However, the sharp increase in transportation demands and vehicle quantities have brought great challenges to existing task assignment methods.

Traditional modeling methods are usually based on simplified constraints and steady-state assumptions, such as mathematical programming (Russell, 2017), graph theory (Xia et al., 2019) and Markov model (Hasan and Ukkusuri, 2017), which are difficult to handle complex and dynamic task assignment problem for vehicles. The rule-based task assignment method can better ensure the real-time decision-making, but the acceptance rate of task assignment and the average return of the system should be further improved. With the storage of vehicles operation data, it is theoretically possible to obtain a decision scheme through existing data learning (Morin et al. 2020). Multi agent can use distributed structure to describe complex and dynamic urban transportation system, so as to reduce the complexity of the system. Reinforcement learning interacts with environment through trial and error, which is suitable for decision-

making problems of the complex dynamic system with large uncertainty and difficult to be solved by traditional methods (Haydari and Yilmaz, 2020). Therefore, the task assignment problem is described as a multi-person multi-stage stochastic game process under cooperative conditions in this paper. A reward-driven decision evaluation method is adopted and the multi-agent reinforcement learning algorithm serves as a solution framework for the problem.

The main works and contributions of this paper include: 1) For task assignment problem, to meet the requirement of immediate response to transportation tasks of users, a stochastic-game-based event-driven task assignment model is developed. It models nodes in transportation network as agents, vehicles at node as agents' resources. Dynamic transportation tasks will trigger the corresponding nodes to make decisions. 2) An extended actor-critic (AC) algorithm is proposed to solve the developed task assignment model and obtain the optimal strategy. This algorithm consists of several actor networks and one centralized critic network. In training process, agents update parameters of actor and critic networks based on experiences of interacting with environment and state value generated by critic network, and achieve ideal synergy. In testing process, agents are able to provide online decision only based on their state. 3) Simulation and comparison experiments was carried out in Didichuxing's open source data (DiDi, 2020), which shows that the proposed model and algorithm for dynamic task assignment of vehicles can significantly improve the acceptance rate of task assignment and the average return of the system. This study can also provide a reference for practical applications.

The rest of paper is organized as follows. Section 2 gives the literature review on the related works. Then in Section 3, we proposed the networked description of urban transportation and developed an event driven task assignment model based on stochastic game. The extended actor-critic algorithm was put forward for model solution in Section 4. Simulation experiments and results analysis are given in Section 5. Finally, the conclusions are summarized in Section 6.

2 Related works

Task assignment problem has always been a hot topic in the fields of enterprise staffing (Bouajaja and Dridi, 2015), factory machine scheduling (Liu et al., 2019), satellite resource scheduling (Gabrel and Vanderpooten, 2002) and transportation (Lin et al., 2001; Srivastava et al., 2008; Glaschenko et al. 2009; Seow et al., 2009; Zhen et al., 2019; Zhang et al., 2018). Transportation task assignment is to reasonably arrange the correspondence between vehicles and tasks, and to propose an immediate task assignment scheme. This problem involves multiple dynamic tasks and limited resources, which is a typical combinatorial optimization problem and also an NP-hard problem. It requires online response to randomly arrived demand, and the information at the time of decision-making is incomplete, including only the current and historical resources and demand information. These features make it difficult to be effectively solved as the general assignment (Chekuri and Khanna, 2005) or knapsack problems (Kleywegt and Papastavrou, 1998). The current literature mainly employs mathematical programming, graph theory, simulation or multi-agent models to solve it.

When the target problem only contains a small-scale task or a single type of resource, the mathematical programming model can be established to obtain the exact solution. Considering the individual and collaborative factors involved, Chen et al. (2009) established a multi-objective optimization model to solve the matching problem between employees and tasks. Some researchers also employed heuristic algorithms to solve complex problems with many constraints, which greatly reduce the computation time and memory consumption. Deng et al. (2016) proposed an accurate algorithm and an approximate algorithm for the matching of staffs

and tasks in the crowdsourcing platform, in which the accurate algorithm is difficult to run because of excessive memory consumption, but the response time of the approximate algorithm is less than millisecond. Abstracted the task allocation problem of unmanned aerial vehicles (UAV) as a collaborative multi-task allocation problem, Jia et al. (2018) developed the mathematical model with kinematic, resource and time constraints, and used the improved genetic algorithm to get the solution of the problem.

The structure of the system can be described intuitively by the node, link and weight in the graph theory model. Gabrel and Vanderpooten (2002) established a graph theory model for the problem of satellite and observation task matching. Further, the shortest path algorithm is used to obtain the task planning scheme to achieve the maximum benefit. Kachroo and Sastry (2016) proposed a travel time function based on traffic density, and established a mathematical programming model to solve the user balance and route allocation schemes by using the node traffic balance in the directed graph with consideration of the intersection time delay.

When the dynamic characteristics cannot be fully expressed by mathematical equations, simulation models can be employed to model the problem. Lin et al. (2001) simulated the freight transportation system in the production logistics by using the combination of the first come first serve rule and the nearest vehicle first rule. Theoretically, the more perfect the actual situation is, the more detailed and accurate the simulation model is, and the more credible the simulation results are. Jorge et al. (2014) confirmed that the mathematical model can get the optimal results, but it needs longer computation time than the simulation model. As for some problems with random and uncertain events, the simulation models can better reflect the effectiveness of the algorithm. However, the modeling and maintenance costs of the simulation models are higher, so it is not suitable for complex systems.

With the advantages of solving large-scale problems, multi-agent systems for task assignment problem have been widely concerned (Srivastava et al., 2008; Seow et al., 2009; Glaschenko et al., 2009; Hao et al., 2013; Lan, 2018). This method essentially enables information sharing between agents through direct or indirect communication to achieve decision sharing. Moreover, some studies have applied multi-agent-based reinforcement learning methods to transportation industry and have achieved good results. A distributed multi-agent deep reinforcement learning method was adopted to solve the problem of controlling traffic signals in a complex urban transportation network, and good results were achieved in terms of optimality and robustness (Chu et al., 2019). Lin et al. (2018) proposed two algorithms based on multi-agent reinforcement learning framework to generate a decision-making scheme for large-scale fleet management of a travel platform. The algorithms can capture supply and demand changes in high-dimensional spaces and formulate corresponding balancing strategies. It is verified in practice that multi-agent systems can significantly improve the utilization of transportation resources. These studies in the context of transportation show that the idea of employing multi-agent reinforcement learning to solve transportation task assignment is feasible.

In short, researches of task assignment problem in many fields are gradually increasing and deepening, and have achieved good results in practical applications. However, there are still some problems such as lack of consideration of random and uncertain factors in practice, and the resulting decision scheme has low flexibility and lag. Especially for the complex and dynamic environment of urban transportation, many algorithms cannot be directly applied. Therefore, in order to improve and solve the above problems, this study proposed a multi-agent reinforcement learning algorithm to solve the problem of transportation task assignment.

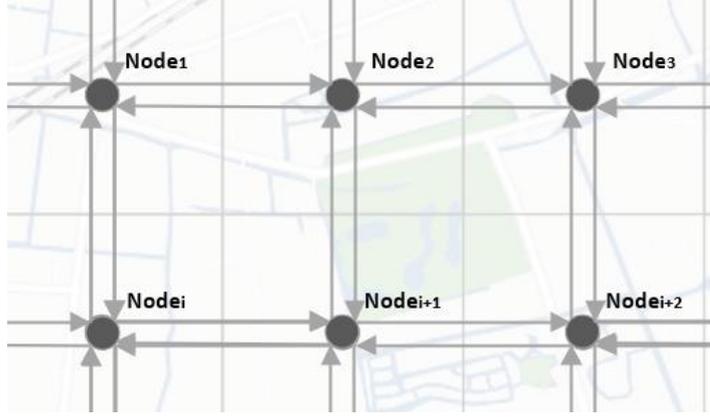


Figure 1: Networked description for urban transportation

3 Event-driven task assignment model based on stochastic game theory

In this section, the networked description of urban transportation is proposed, and an event driven task assignment model based on stochastic game is developed.

3.1 Networked Description of Urban Transportation

Based on the idea of graph theory, the complex urban transportation system is abstracted as a complex network $G = (N, E)$ composed of nodes and edges (see Figure 1). $N = \{Node_1, Node_2, \dots, Node_n\}$ is the set of nodes in the complex network, which represent various areas of urban roads. $E = \{Edge_{12}, Edge_{21}, \dots, Edge_{ij}\}$ is the set of edges in the complex network. There are two edges $Edge_{ij}$ and $Edge_{ji}$ connected between any two adjacent nodes $Node_i$ and $Node_j$. In our opinion, any known urban transportation system can be described by G .

Vehicles and tasks in the transportation network are denoted by V and T , where $V = \{Vehicle_1, Vehicle_2, \dots\}$ is the set of vehicles, and $T = \{Task_1, Task_2, \dots\}$ is the set of tasks. We defined c_{it} as the total vehicle resource at $Node_i$ at time t , $l_{i,t}$ as the total transport task for $Node_i$ at time t . The service period is usually divided into days or months, which is expressed as P . In order to describe the dynamic changes in the environment and resources, time is discretized, and the service period of the vehicle between any two adjacent nodes is taken as the time interval Δt .

Before developing task assignment model, we made the following assumptions based on the networked description for urban transportation:

1) Modeling objects are moments and places where demand is greater than supply. Based on analysis of real scenarios, when demand is less than supply or supply and demand are balanced, as long as any demand arrives, timely response can ensure that the global benefit is maximized. In that case, no task assignment and evaluation are required.

2) Each period in service cycle is the assignment period of the transportation task, $P = [P^{start}, P^{end}]$ where P^{start} is the start time of the round of assignment, P^{end} is the end time of the round of assignment, $\Delta t = P^{end} - P^{start}$ is the time interval.

3) Vehicle resources of node are updated before start time P^{start} , which includes: the remaining vehicles of the node in the previous period, the vehicles that arrived from other nodes in the previous period, and the vehicles that completed the transportation task to reach the destination node.

- 4) In the same assignment period, except for assignment decisions, the number of vehicles at a node will not increase or decrease due to external factors. The number of vehicles at a node is the maximum number of tasks that the node can accept during this period.
- 5) Transport tasks are represented as $task = \{No^{task}, t^e, t^w, t^d, node^{dep}, node^{dest}, v, m\}$, where No^{task} is the number of tasks; t^a , t^w and t^d denote the task assignment, waiting and delivery time, respectively; $node^{dep}$ and $node^{dest}$ are places of departure and destination, respectively; v is task value and m is the order in which tasks arrive.
- 6) For tasks that are not accepted during the task assignment period, if there is a waiting time $t^w \neq 0$, the assignment request can be re-initiated in multiple assignment periods of $t^a + t^d$, and it has higher priority in new assignment period, which means $m^{old} < m^{new}$.

3.2 Stochastic game and model development

Stochastic game. Multi-agent reinforcement learning has the characteristics of multi-stage of the Markov decision process, and also has the characteristics of multi-participant of matrix games, so it is usually expressed by a stochastic game that combines the two. Stochastic game is a type of dynamic game with state probability transition, which is performed by one or more participants. It can be defined as:

$$SG = (n, S, A, P, R_i) \quad (1)$$

where n is the number of agents; S is the state set of the environment; A_i refers to action set that agent i can choose; P represents the state transition probability; R_i is the agent's return function. In this process, multiple agents make a choice of actions, and the next state and reward of the environment is determined by the joint actions of multiple agents (see Figure 2).

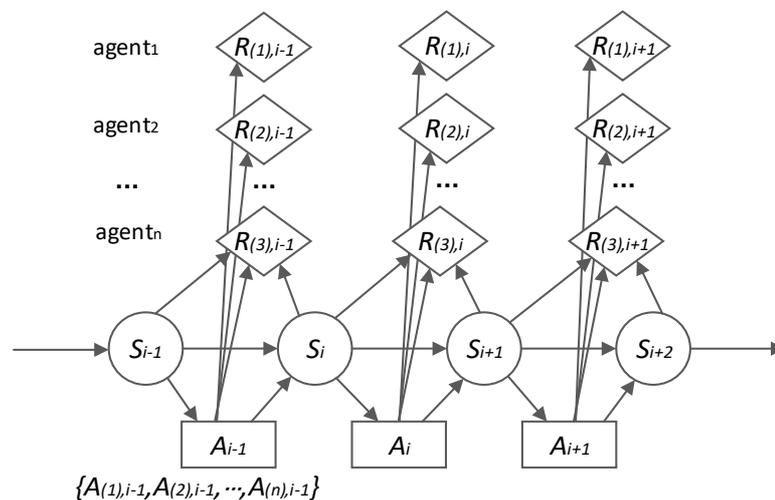


Figure 2: Stochastic game

Stochastic games are aimed at solving the Nash equilibrium, but under normal conditions, the transfer function and return function are unknown. In reinforcement learning, the agent learns the equilibrium strategy through interaction with the environment, and uses the rationality and convergence to evaluate algorithm performance (Bowling and Veloso, 2002).

Agent. Each node in the transportation network is considered as an agent. Without considering factors such as driver's historical order acceptance rate and preferences, the vehicles are no difference in the same or similar locations. Therefore, each node has two states: demand

vehicles or supply vehicles, which also denotes agent states. Vehicles are the resource owned by nodes, that is, attributes of agents. In practice, there is a one-to-one assignment relationship between the transportation task and the vehicle. If each vehicle is assigned to a transportation task, most of the other joint actions are invalid. Compared with considering vehicles as agents in the literature (Gupta et al. 2017), our agent setting method can greatly reduce the number of agents, and further reduce the environment's joint action space and calculations.

State. When task arrives, the task's destination and the estimated value can be observed by the node. The resource of other nodes has little influence on the decision of the vehicles in this node, thus only the remaining vehicle resources of this node are considered. The environmental information observed by each agent can be defined as the resource remaining, task information and time information of the node where the vehicle is located:

$$s_i = \{c_i^{remain}, s_i^{task}, s_i^{time}\} \quad (2)$$

where c_i^{remain} is the remaining resources of current node and s_i^{time} is the current assignment time. The task that arrives can be expressed as,

$$task_i = \{i, t^e, t^w, t^d, node^{dep}, node^{dest}, v_i, m\} \quad (3)$$

where the state of the task can be represented as $s_i^{task} = \{node^{dest}, v_i\}$.

Action. For any task k that arrives at node i , its departure node and its neighboring nodes can choose whether to accept the task,

$$a_{i,k} = \{0,1\} \quad (4)$$

where $a_{i,k} = 0$ denotes that the task is rejected and $a_{i,k} = 1$ denotes that the task is accepted.

Reward. The rewards obtained from the interactive feedback between nodes and the environment are determined by the node state and actions simultaneously. When the $task_k = \{k, t^e, t^w, t^d, node^{dep}, node^{dest}, v_i, m\}$ arrives at t , the reward received by the node is defined as,

$$r_{i,k} = \begin{cases} 0, & \text{when node } i \text{ rejects task } k \\ \alpha \frac{v_i}{\sum a_{i,k}} + \beta c_j^{remain}, & \text{when node } i \text{ accepts task } k \end{cases} \quad (5)$$

where v_i is task value and $\sum a_{i,k}$ is the number of nodes that choose to accept the task. When more nodes choose to accept the task, the nodes can get less rewards. c_j^{remain} is the remaining resources of the current node. When there are more remaining resources j , the greater the reward that the node can get, the more inclined it is to accept the task. α and β are normalized coefficients for task value and resource consumption, which is to eliminate the difference in feature vector values of different dimensions.

State probability transition. The vehicle resource distribution and node location information during the service period are known, but the specific information of the next arrival task is unknown. And the environment condition will refresh between periods, so that the vehicle

resource distribution changes on each node and the state transition probability function is unknown.

End time. For the entire assignment process, task assignment is terminated when the service cycle ends. In a certain assignment period, when the vehicle resources of each node in the transportation network run out, the next assignment period is started.

$$\begin{cases} t^{end} = T \\ t^{curr} = t^{curr} + \Delta t, \text{ if } \sum_{node_i \in Node} c_{i,t^{curr}} = 0 \end{cases} \quad (6)$$

where t^{curr} refers to the current time of the environment and $\sum_{node_i \in Node} c_{i,t^{curr}}$ is the total number of vehicle resources in the transportation network during the t^{curr} period.

4 Extended actor-critic algorithm for model solution

The AC algorithm (Konda and Tsitsiklis, 2000; Bhatnagar et al., 2008; Babaeizadeh et al., 2016) is the basic framework we adopt, which combines value function-based and policy gradient-based methods, improves the limit of the state space dimension in the value function-based method, and solves the randomness of the environment that causes the estimated policy gradient to have a large variance in multiple samplings. The framework consists of two networks, one is the actor network $\pi(s, a, \theta)$, which is used to optimize agent strategies; the other is the critic network $\hat{q}(s, a, \omega)$, which is used to estimate the value function. Parameters of the neural network are θ and ω , respectively. Based on critic's evaluation for the action taken, actor will adjust its strategy, and critic will update the value function based on experience and rewards.

4.1 Network Structure

In the extended AC framework, we establish different actor networks for different agents, which can maintain its own network parameters. In actual situations, there is a difference in the probability distribution of tasks arriving at different locations in the city. For example, the tasks at the center of the city have a short distance and a short time, and tasks at the edge of the city may take longer and be more valuable higher. If a network is simply described by shared parameters, the differences between nodes cannot be reflected, which may cause problems such as the difficulty in convergence of results. Therefore, we proposed a centralized training and distributed execution structure. During the training process, each agent learns strategies from observations and actions of its own environment. A centralized critic network uses the observation status of each node as input, and updates the rewards obtained by the actor's action feedback based on the environment. In this process, centralized training can make the strategies of each agent achieve tacit coordination, while decentralized execution can extract the local strategies of each agent from the global strategy, thereby achieving the purpose of task assignment.

Figure 3 shows the distributed network structure used in this study. There are two parts, multiple networks for executing strategies and a centralized value function network. Strategy network and value function network in the figure are both multi-layer feedforward neural networks with three layers.

4.2 Network Training

Actor's policy gradient is calculated by,

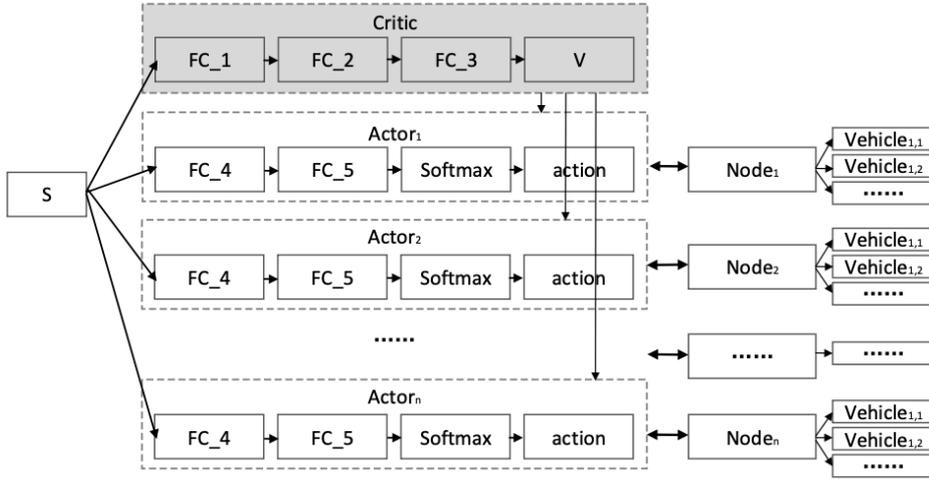


Figure 3: Extended AC decision model framework

Table 1: Extended AC algorithm

input: environment, number of iterations N , period T , number of nodes n , state space dimension, action space dimension, step size α, β , attenuation factor γ , exploration rate ε , critic network structure and actor network structure
output: actor network parameters $\theta_1, \theta_2, \dots, \theta_n$, critic network parameters ω

Initialize network parameters

for i from 1 to N do

 Initialize the environment and get the initial state s_0

 for t from 1 to T do

$j = 0$

 while there are tasks and vehicle resources left at target node

 for k from 1 to n do

 use s_t in the network as input, output action $a_{t,k}$

 perform actions to get feedback r_t and next state s_{t+1}

 end for

 calculate dominance function and target critic network value

 function

$j = j + 1$

 for m from j to 1 do

 Actor network parameter update

 Critic network parameter update

 end for

 end for

end for

$$\nabla J(\theta) = E_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s_t, a_t) V(s_t, \omega)] = E_{\pi_\theta} [\nabla_\theta \log \pi_\theta(s_t, a_t) A(s_t, t, \omega)] \quad (7)$$

Advantage function A is used as the evaluation point of critic network, which can be defined as the difference between the action value function and the state value function, and replaced by its unbiased estimate.

$$A(s,t) = r + \gamma V(s_{t+1}) - V(s) \quad (8)$$

Critic network loss is the squared loss of actual state value and estimated state value, and its parameters are updated using time difference (TD).

$$\begin{aligned} \min \sum (V_{\pi}(s_{t+1}, \omega') - V(s_t, \omega))^2 \\ V_{\pi}(s_{t+1}, \omega') = \sum \pi(s_t, a_t)(r_{t+1} + \gamma V(s_{t+1}, \omega')) \end{aligned} \quad (9)$$

Whenever a new task arrives, vehicles at the same node will accept the same decision, that is, intelligent node will give a unified decision of vehicles at that node, and select one of the vehicles to complete the real matching action. This method can reduce the task contradictions between matching decisions. In addition, when multiple nodes are involved in task matching, there may still be conflicts between the actions given by the nodes. In order to meet the constraint of the task's uniqueness, the state value generated by the centralized evaluation network is used as the basis for the final action selection for task coordination between nodes. The pseudocode is shown in [Table 1](#).

5 Experiment

5.1 Data and Simulation Environment

Didichuxing's open source data ([DiDi, 2020](#)) is used to verify the effectiveness of the proposed method. Some data samples are shown in [Table 2](#). By analyzing and visualizing the data, it can be seen that the attributes of the task have different characteristics in different periods, such as 7: 30-7: 40 and 19: 50-20: 00, as shown in [Figures 4, 5](#), respectively. The tasks submitted in the two periods are divided according to the places of departure and destination. The number of tasks contained in each place can be seen from the figure. The place of departure is more scattered, and the place of destination is relatively concentrated for the period 7: 30-7: 40, while the period 19: 50-20: 00 is the opposite. The results of this analysis are also consistent with actual life experiences.

Table 2: Data samples

Orders number	mjiwdgk	f78cfb7e	5c33acbf	...
Start billing time	1501581031	1477963587	1477965143	...
End billing time	1501582195	1477965143	1477959461	...
Longitude of departure position	104.11225	104.05412	104.07139	...
Latitude of departure position	30.66703	30.67206	30.71631	...
Longitude of destination position	104.07403	104.06614	104.05733	...
Latitude of destination position	30.686300	30.709336	30.699250	...

The proposed method uses a distributed network structure with high complexity. In order to reduce training costs and time, we only considered a part of nodes in the urban transportation network. In this experiment, five nodes were selected as the modeling objects. The total number of tasks and the total number of vehicle resources for these selected regional nodes are shown in [Figure 6](#). The average order acceptance rate of the nodes is about 82.866%, which means the demand for vehicle resources exceeds the supply for a long time.

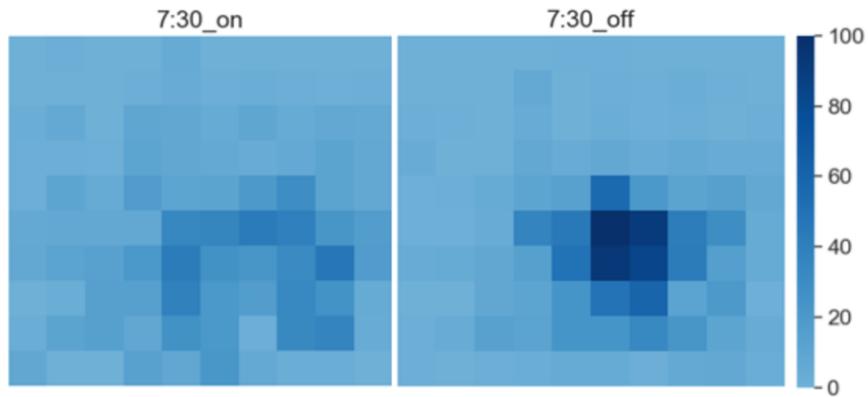


Figure 4: Distribution of orders' departure and arrival in 7:30-7:40

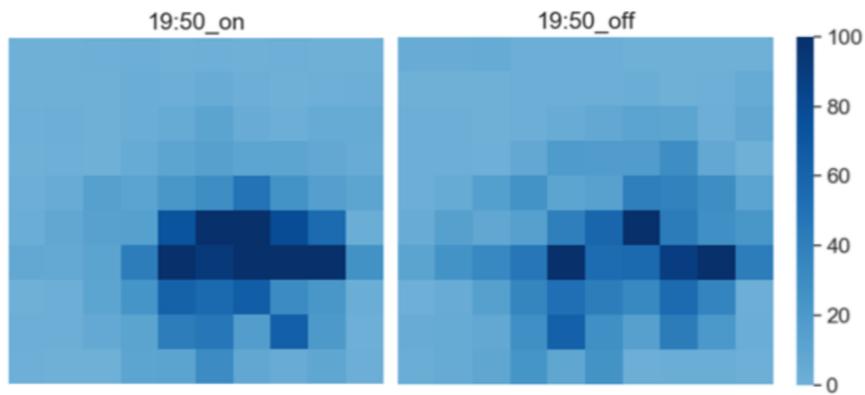


Figure 5: Distribution of orders' departure and arrival in 19:50-20:00

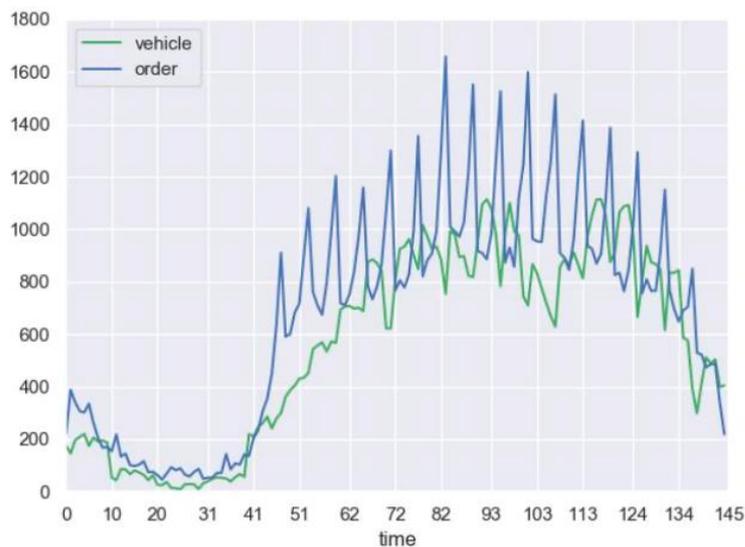


Figure 6: Number of orders and vehicles in local areas

5.2 Result Analysis

Task acceptance rate and profit rate are used to evaluate the performance of the method. Task acceptance rate is the ratio of the number of tasks accepted to the total number of tasks, and profit rate is the ratio of the total value of accepted tasks to the total task value.

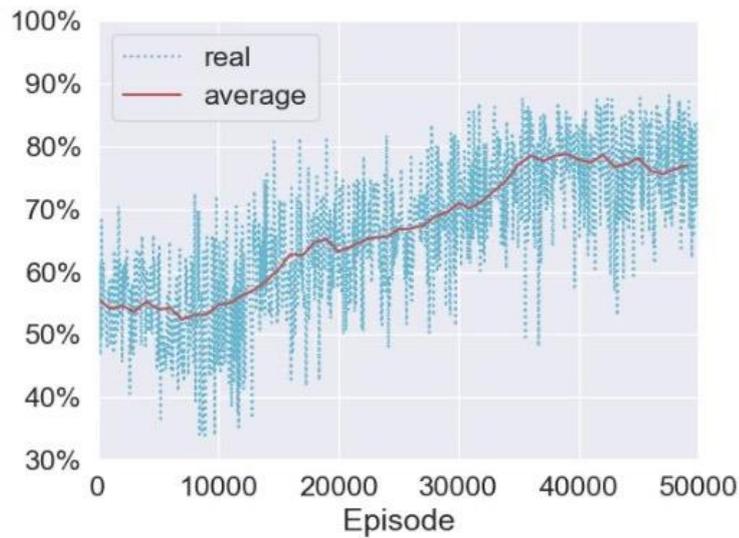


Figure 7: Task acceptance rate change with training episodes

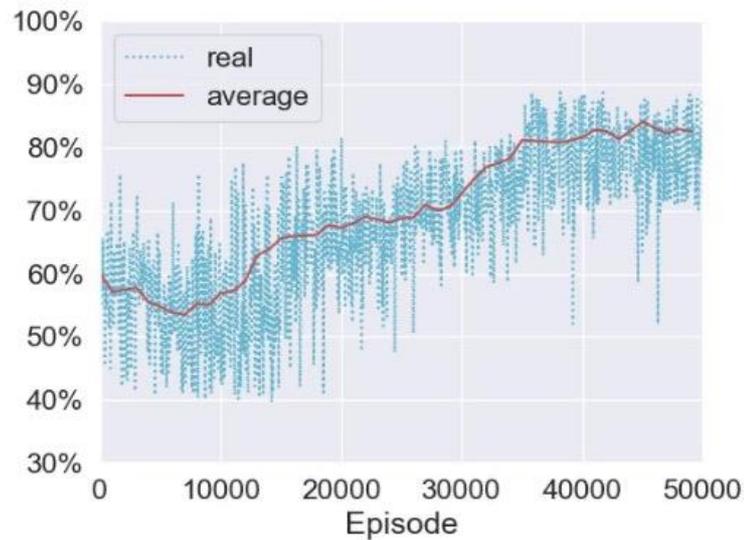


Figure 8: Profit rate change with training episodes

Figures 7 and 8 show the changes in the task acceptance rate and profit rate of the proposed method with training rounds, where blue is the true value and red is the average. The trends in the two figures are basically consistent. At the beginning of training, the agent belongs to the tentative exploration stage, and the task acceptance rate and profit rate have both declined slightly, but then gradually increased. It can be seen that the task acceptance rate and profit rate gradually stabilized after training to about 35,000 rounds, when the model gradually converges.

Figure 9 shows the number of remaining vehicles in each round during the training process. A large amount of vehicle resources was idle during the initial training period. By learning the acceptance and rejection strategy for transportation tasks, the wasted vehicle resources of each modeling node are reduced gradually.

First-come-first-served (FCFS) task assignment method and contract network are used for comparing with the proposed method. FCFS method means that if a transport task arrives and the task departure node still has idle vehicles remaining, the task is assigned to it. If there is no idle vehicle remaining at the node, the transportation task is assigned to the neighboring node

with idle vehicles. The contract net algorithm (CNA) is appropriately modified to fit the context of this paper (Hu et al., 2019).

Table 3 compares the experimental results of different algorithms with the proposed method. As can be seen from the data in this table, the task acceptance rate of each algorithm is similar, but our method performs better in profit rate, which means that it is effective and can make the vehicle obtain greater benefits.

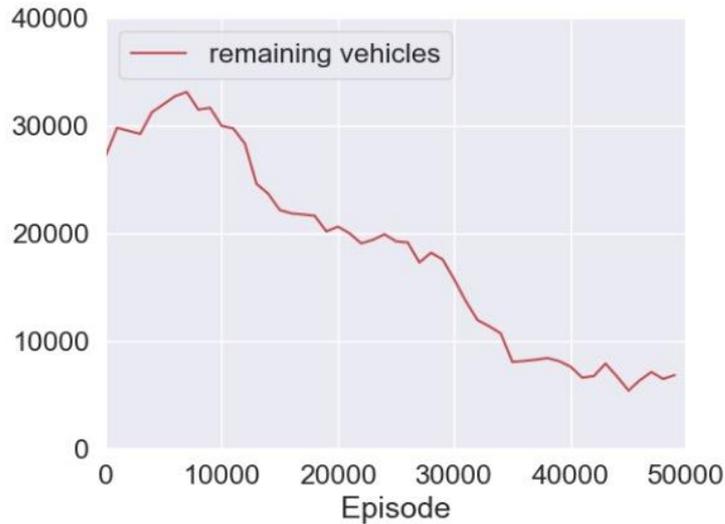


Figure 9: Number of remaining vehicles change with training episodes

Table 3: Comparison of task acceptance rate and profit rate of different algorithms

Algorithms	task acceptance rate	profit rate
FCFS	89.373%	87.068%
CNA	89.565%	88.203%
Proposed method	89.555%	89.159%

6 Conclusion

A reasonable and efficient task assignment method is the direct means to improve the revenue. This paper proposed a dynamic task assignment method for vehicles in urban transportation based on multi-agent reinforcement learning. Aiming at the problem of unreasonable task assignment due to greedy choice, an event-driven random game model was developed to describe the task assignment problem of vehicles. An extended actor-critic (AC) algorithm is proposed for model solution. The distributed network structure is used to construct a learning framework with the positions of various nodes as the decision-making subject in the urban transportation network. By comparing with the mainstream task assignment methods, our method can make vehicle operators achieve higher revenues while ensuring immediate response to transportation tasks.

Since the adopted framework involves the parallel computation of multiple neural networks and takes a long time for training and parameter optimization, the proposed method still has some shortcomings. In the subsequent research, the framework structure or mapping relationship can be further optimized to reduce its complexity and thus have more practical value.

Acknowledgment

The authors would like to acknowledge financial supports of the National Natural Science Foundation of China (No. 51775348), and the National Key Research and Development Program of China (No. 2019YFB1704401).

References

- Babaeizadeh, M., Frosio, I., Tyree, S., Clemons, J., Kautz, J. (2016). Reinforcement learning through asynchronous advantage actor-critic on a gpu. arXiv preprint arXiv:1611.06256.
- Ballot, E., Montreuil, B., Meller, R. (2014): The physical internet.
- Bhatnagar, S., Ghavamzadeh, M., Lee, M., Sutton, R. S. (2008). Incremental natural actor-critic algorithms. *Advances in neural information processing systems*, 105-112.
- Bouajaja, S., Dridi, N. (2015). Research on the optimal parameters of ACO algorithm for a human resource allocation problem. *2015 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, 60-65, doi: 10.1109/SOLI.2015.7367412.
- Bowling, M., Veloso, M. (2002). Multiagent learning using a variable learning rate. *Artificial Intelligence*, v136, no2, 215-250.
- Chekuri, C., Khanna, S. (2005). A polynomial time approximation scheme for the multiple knapsack problem. *SIAM Journal on Computing*, v35, no3, 713-728.
- Chen, X., Fan, Z. P., Li, Y. H. (2009). Matching Problem of Employee and Task Based on Individual and Cooperative Factors. *Industrial Engineering and Management*, v14, no2, 120-124. (in Chinese).
- Chu, T., Wang, J., Codecà, L., Li, Z. (2019). Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, v21, no3, 1086-1095.
- Deng, D., Shahabi, C., Demiryurek, U., Zhu, L. (2016). Task selection in spatial crowdsourcing from worker's perspective. *GeoInformatica*, vol20, no3, 529-568.
- DiDi. (2020). GAIA open dataset. <https://gaia.didichuxing.com>.
- Gabrel, V., Vanderpooten, D. (2002). Enumeration and interactive selection of efficient paths in a multiple criteria graph for scheduling an earth observing satellite. *European Journal of Operational Research*, v139, no3, 533-542.
- Glaschenko, A., Ivaschenko, A., Rzevski, G., Skobelev, P. (2009). Multi-agent real time scheduling system for taxi companies. *8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, 29-36.
- Gupta, J. K., Egorov, M., Kochenderfer, M. (2017). Cooperative multi-agent control using deep reinforcement learning. *International Conference on Autonomous Agents and Multiagent Systems* Springer, Cham, v10642, 66-83.
- Hao, H., Jiang, W., Li, Y., Yuan, Z. (2013). Research on agile satellite dynamic mission planning based on multi-agent. *Journal of National University of Defense Technology*, v35, no1, 53-59. (in Chinese).
- Hasan, S., Ukkusuri, S. V. (2017). Reconstructing activity location sequences from incomplete check-in data: a semi-Markov continuous-time Bayesian network model. *IEEE Transactions on Intelligent Transportation Systems*, v19, no3, 687-698.
- Haydari, A., Yilmaz, Y. (2020). Deep reinforcement learning for intelligent transportation systems: a survey. *IEEE Transactions on Intelligent Transportation Systems*, doi: 10.1109/TITS.2020.3008612.

- Hu, Y., Li, C., Zhang, K., Fu, Y. (2019). Task allocation based on modified contract net protocol under generalized cluster. *Journal of Computational Methods in Sciences and Engineering*, v19, no4, 969-988.
- Jia, Z., Yu, J., Ai, X., Xu, X., Yang, D. (2018). Cooperative multiple task assignment problem with stochastic velocities and time windows for heterogeneous unmanned aerial vehicles using a genetic algorithm. *Aerospace Science and Technology*, v76, 112-125.
- Jorge, D., Correia, G., H., A., Barnhart, C. (2014). Comparing optimal relocation operations with simulated relocation policies in one-way carsharing systems. *IEEE Transactions on Intelligent Transportation Systems*, v15, no4, 1667-1675.
- Kachroo, P., Sastry, S. (2016). Traffic assignment using a density-based travel-time function for intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, v17, no5, 1438-1447.
- Kaffash, S., Nguyen, A. T., Zhu, J. (2020). Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis. *International Journal of Production Economics*, v231, no107868, 1-15.
- Kleywegt, A. J., Papastavrou, J. D. (1998). The dynamic and stochastic knapsack problem. *Operations research*, v46, no1, 17-35.
- Konda, V., R., Tsitsiklis, J., N. (2000). Actor-critic algorithms. *Advances in neural information processing systems*.
- Kubek, D., Więcek, P. (2019). An integrated multi-layer decision-making framework in the physical internet concept for the city logistics. *Transportation Research Procedia*, v39, 221-230.
- Lan, C. (2018). Research on multi-task rapid scheduling technology for satellite networks. M. S. thesis, Xidian University, China. (in Chinese).
- Lin, J., T., Wang, F., K., Yen, P., Y. (2001). Simulation analysis of dispatching rules for an automated interbay material handling system in wafer fab. *International Journal of Production Research*, v39, no6, 1221-1238.
- Liu, J., L., Wang, L., C., Chu, P., C. (2019). Development of a cloud-based advanced planning and scheduling system for automotive parts manufacturing industry. *Procedia Manufacturing*, v38, 1532-1539.
- Lin, K., Zhao, R., Xu, Z., Zhou, J. (2018). Efficient large-scale fleet management via multi-agent deep reinforcement learning. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1774-1783.
- Morin, M., Gaudreault, J., Brotherton, E., Paradis, F., Rolland, A., Wery, J., Laviolette, F. (2020). Machine learning-based models of sawmills for better wood allocation planning. *International Journal of Production Economics*, v222, no107508, 1-10.
- Russell, R. A. (2017). Mathematical programming heuristics for the production routing problem. *International Journal of Production Economics*, v193, 40-49.
- Seow, K. T., Dang, N. H., Lee, D. H. (2009). A collaborative multiagent taxi-dispatch system. *IEEE Transactions on Automation science and engineering*, v7, no3, 607-616.
- Srivastava, S. C., Choudhary, A. K., Kumar, S., Tiwari, M. K. (2008). Development of an intelligent agent-based AGV controller for a flexible manufacturing system. *The International Journal of Advanced Manufacturing Technology*, v36, no7-8, 780.
- Xia, F., Wang, J., Kong, X., Zhang, D., Wang, Z. (2019). Ranking station importance with human mobility patterns using subway network datasets. *IEEE Transactions on Intelligent Transportation Systems*, v21, no7, 2840-2852.

- Zhang, Y. H., Gong, Y. J., Chen, W. N., Gu, T. L., Yuan, H. Q., Zhang, J. (2018). A dual-colony ant algorithm for the receiving and shipping door assignments in cross-docks. *IEEE Transactions on Intelligent Transportation Systems*, v20, no7, 2523-2539.
- Zhen, L., Yu, S., Wang, S., Sun, Z. (2019). Scheduling quay cranes and yard trucks for unloading operations in container ports. *Annals of Operations Research*, v273, no1, 455-478.
- Zhong, R. Y., Xu, C., Chen, C., Huang, G. Q. (2017). Big data analytics for physical internet-based intelligent manufacturing shop floors. *International journal of production research*, v55, no9, 2610-2621.

A Two-Stage Production Planning Model for Perishable Products Under Uncertainty

Kin Keung Lai¹, Ming Wang^{1*}

¹ College of Economics, Shenzhen University, Shenzhen, China

Abstract: - This study addresses the production planning problem for perishable products, in which the cost and shortage of products are minimized subject to a set of constraints such as warehouse space, labor working time and machine time. Using the concept of postponement, the production process for perishable products is differentiated into two phases to better utilize the resources. A two-stage stochastic programming with recourse model is developed to determine the production loading plan with uncertain demand and parameters. A set of data from a toy company shows the benefits of the postponement strategy: these include lower total cost and higher utilization of resources. Comparative analysis of solutions with and without postponement strategies is performed.

Key-Words: - Production planning, Stochastic Programming, Modeling, Perishable

1 Introduction

Items like dairy products, medical products and chemical products cannot be stored for a long time because they rot or can no longer be used. For other items such as computers and mobile phones, sale volumes drop dramatically when a new generation is introduced. Seasonal products like high fashion apparel, Christmas gifts and calendars are sold only below full price after a day or a season. These products are regarded as perishable products. Controlling the inventory of perishable products is crucial. On one hand, the demand for perishable products is time-sensitive. This means that the demand dramatically increases as the day approaches the end of life-cycle, such as Christmas Day. On the other hand, a shortage of perishable products while the products are saleable may result in significant loss of revenue because the perishable products cannot be profitable after a certain day. For instance, in manufacturing industries, people want to buy Christmas gifts in or before December only. However, there is little research that addresses aggregate production planning for perishable products. In order to deal with the production planning under limited resources while facing a dramatic growth in demand, in this study we employ a postponement strategy in production planning for perishable products. Postponement in production planning refers to common intermediate products being manufactured in a first phase, and, according to the differentiating options such as colors, sizes and types, production line activities such as dyeing, compounding, final assembling, packaging and so on are postponed to a second phase until customer orders received [1], [2], [7]. Hence, with a postponement strategy, we determine (1) how many finished products should be produced from raw materials directly (direct production), (2) how many semi-finished products should be produced from raw materials (master production), and (3) how many finished products

*Corresponding author:

Email address : wmtroy@outlook.com

should be produced from semi-finished products (final assembly) so that the resources can be better utilized to meet the dramatic growth in demand. A well-known real-life postponement example is the redesign of the European DeskJet Printer line by Hewlett Packard, as illustrated by Lee and Billington [5].

However, no research is found to solve aggregate production planning of perishable products under an uncertain environment. The purpose of this study is to develop a stochastic programming model to optimize the production planning problem for perishable products; from this the optimal production plan and workforce level for a medium-term planning horizon is determined with the minimal total costs consisting of the production cost, setup cost, labor cost, inventory cost, hiring cost and lay-off cost, and penalty cost associated with under-fulfillment of realized demand under different economic growth scenarios.

2 Problem Formulation

One of the widely-used formulations for decision making under uncertainty is stochastic programming with recourse. The basic idea of this modeling approach is to formulate the problem in a two-stage setting. In the first stage, a decision is made based on the deterministic parameters. When the uncertainty is realized, a corrective recourse action is then made at the second stage. The objective of two-stage stochastic program is to minimize the total costs associated with the first stage decision and the expected future recourse costs at the second stage. The incorporation of the expected future recourse costs provides a proactive approach to tackle the future uncertainty at the beginning of modeling. For detail discussion of the approach of two-stage stochastic programming with recourse, the reader is referred to Dantzig [3], Kall and Wallace [4] and Ruszczyński and Shapiro [6].

In this study, the aggregate production planning problem for perishable products faced by a toy company in Hong Kong is investigated. For cost effectiveness, the decision makers have to determine the quantity of product i , $i = 1, 2, \dots, n$, manufactured over each period of time t , $t = 1, 2, \dots, T$, to fulfill market demands under different scenarios s , $s = 1, 2, \dots, S$. The production loading plan consists of: (1) the quantity of finished products to be produced from raw materials directly (direct production), (2) the quantity of semi-finished products to be produced from raw materials (master production), and (3) the quantity of finished products to be produced from semi-finished products (final assembly) in each period of time.

2.1 Notation

Parameters:

First-stage parameters:

C_{KXi} : the setup cost for producing finished product i from raw materials

C_{KYi} : the setup cost for producing semi-finished product i from raw materials

C_{KZi} : the setup cost for producing finished product i from semi-finished products

C_{Wt} : the labor cost in period t

C_{Ht} : the cost to hire one worker in period t

C_{Lt} : the cost to lay-off one worker in period t

\overline{W}_t : the maximum number of workers available in period t

Second-stage parameters:

D_{it}^s : the forecast demand for product i in period t under scenario s

C_{PXi}^S : the regular-time unit production cost to produce one unit of finished product i from raw materials under scenario s

C_{PYi}^S : the regular-time unit production cost to produce one unit of semi-finished product i from raw materials under scenario s

C_{PZi}^S : the regular-time unit production cost to produce one unit of finished product i from semi-finished products under scenario s

C_{OXi}^S : the overtime unit production cost to produce one unit of finished product i from raw materials under scenario s

C_{OYi}^S : the overtime unit production cost to produce one unit of semi-finished product i from raw materials under scenario s

C_{OZi}^S : the overtime unit production cost to produce one unit of finished product i from semi-finished products under scenario s

$C_{\alpha i}^S$: the inventory holding cost for one unit of finished product i under scenario s

$C_{\beta i}^S$: the inventory holding cost for one unit of semi-finished product i under scenario s

$C_{U i}^S$: the cost of under-fulfillment for one unit of finished product i under scenario s

a_{Xi} : the man hours required to produce one unit of finished product i from raw materials

a_{Yi} : the man hours required to produce one unit of semi-finished product i from raw materials

a_{Zi} : the man hours required to produce one unit of finished product i from semi-finished products

b_{Xi} : the machining time required to produce one unit of finished product i from raw materials

b_{Yi} : the machining time required to produce one unit of semi-finished product i from raw materials

b_{Zi} : the machining time required to produce one unit of finished product i from semi-finished products

δ : the regular working hours of labor in each period

λ_t^W : the fraction of regular workforce available

for over-time in period t

λ_t^M : the fraction of regular machine capacity available for over-time use in period t

M_t : the maximum regular time machine capacity in period t

$v_{\alpha i}$: the space occupied by one unit of finished product i

$v_{\beta i}$: the space occupied by one unit of semi-finished product i

\bar{I}_t : the storage space limitation in period t

Decision variables:

First-stage decision variables:

K_{Xit} : the indicator for producing finished product i from raw materials in period t (if $K_{Xit} = 1$, then $P_{Xit} > 0$; if $K_{Xit} = 0$, then $P_{Xit} = 0$)

K_{Yit} : the indicator for producing finished product i from raw materials in period t (if $K_{Yit} = 1$, then $P_{Yit} > 0$; if $K_{Yit} = 0$, then $P_{Yit} = 0$)

K_{Zit} : the indicator for producing finished product i from raw materials in period t (if $K_{Zit} = 1$, then $P_{Zit} > 0$; if $K_{Zit} = 0$, then $P_{Zit} = 0$)

H_t : the number of workers hired in period t

L_t : the number of workers laid-off in period t

W_t : the number of workers in period t

Second-stage decision variables:

P_{Xit} : the number of finished products i produced from raw materials during regular time in period t

P_{Yit} : the number of semi-finished products i produced from raw materials during regular time in period t

P_{Zit} : the number of finished products i produced from semi-finished products during regular time in period t

O_{Xit} : the number of finished products i produced from raw materials during overtime in period t

O_{Yit} : the number of semi-finished products i produced from raw materials during overtime in period t

O_{Zit} : the number of finished products i produced from semi-finished products during overtime in period t

$I_{\alpha it}^s$: the inventory level of finished product i in period t under scenario s

$I_{\beta it}^s$: the inventory level of semi-finished product i in period t under scenario s

U_{it}^s : the under-fulfillment of finished product i in period t under scenario s

2.2 Objective function

The objective function at the first stage:

$$\text{Min} \sum_{t=1}^T \sum_{i=1}^n (C_{KXi}K_{Xit} + C_{KYi}K_{Yit} + C_{KZi}K_{Zit}) + \sum_{t=1}^T C_{Wt}W_t + \sum_{t=1}^T (C_{Ht}H_t + C_{Lt}HL_t) \quad (1)$$

The first term in expression (1) is the setup cost. The second term is the labor cost, which is associated with regular-time workers. The last term is total hiring and laying-off cost associated with changes in the workforce level.

The objective function at the second stage:

$$\text{Min} \sum_{s=1}^S P_s \left[\sum_{t=1}^T \sum_{i=1}^n (C_{PXi}^s P_{Xit} + C_{PYi}^s P_{Yit} + C_{PZi}^s P_{Zit}) + \sum_{t=1}^T \sum_{i=1}^n (C_{OXi}^s O_{Xit} + C_{OYi}^s O_{Yit} + C_{OZi}^s O_{Zit}) \right. \\ \left. + \sum_{t=1}^T \sum_{i=1}^n (C_{\alpha i}^s I_{\alpha it}^s + C_{\beta i}^s I_{\beta it}^s + C_{U_i}^s U_{it}^s) \right] \quad (2)$$

The first term in expression (2) is the regular-time production cost, which comprises associated direct production, master production and final assembly. The second term is the over-time production cost, which is associated with direct production, master production and final assembly. The third term is the inventory cost associated with the storage of units of finished products and semi-finished products in warehouses for a period of time. The last term is the penalty cost associated with under-fulfillment of demand.

2.3 Constraints

The constraints at the first stage:

$$W_t = W_{t-1} + H_t - L_t, t = 1, 2, \dots, T \quad (3)$$

$$W_t \leq \overline{W}_t, t = 1, 2, \dots, T \quad (4)$$

$$W_t, H_t, L_t \geq 0, i = 1, 2, \dots, n, t = 1, 2, \dots, T \quad (5)$$

$$K_{Xit}, K_{Yit}, K_{Zit} = \{0,1\}, i = 1,2, \dots, n, t = 1,2, \dots, T \quad (6)$$

Constraint (3) ensures that the available workforce in any period equals the workforce from the previous period plus any change in workforce level during the current period. The change in workforce level may be due to either hiring extra workers or laying-off redundant workers. It is noted that $H_t * L_t = 0$ because either the net hiring or the net laying-off of workers takes place over a period, but not both. Constraint (4) ensures the upper-bounds of change in workforce level over a period are provided. Constraint (5) ensures that all decision variables are non-negative. Boolean constraints (6) are used for the setup indications of the production activities.

The constraints at the second stage:

$$I_{ait}^s - U_{it}^s = I_{ait-1}^s + P_{Xit} + O_{Xit} + P_{Zit} + O_{Zit} - D_{it}^s, i = 1,2, \dots, n, t = 1,2, \dots, T \quad (7)$$

$$I_{\beta it}^s = I_{\beta it-1}^s + P_{Yit} + O_{Yit} - P_{Zit} - O_{Zit}, i = 1,2, \dots, n, t = 1,2, \dots, T \quad (8)$$

$$\sum_{i=1}^n (v_{\alpha i} I_{ait}^s + v_{\beta i} I_{\beta it}^s) \leq \bar{I}_t, t = 1,2, \dots, T \quad (9)$$

$$\sum_{i=1}^n (a_{Xi} P_{Xit} + a_{Yi} P_{Yit} + a_{Zi} P_{Zit}) \leq \delta W_t, t = 1,2, \dots, T \quad (10)$$

$$\sum_{i=1}^n (a_{Xi} O_{Xit} + a_{Yi} O_{Yit} + a_{Zi} O_{Zit}) \leq \lambda_t^M \delta W_t, t = 1,2, \dots, T \quad (11)$$

$$\sum_{i=1}^n (b_{Xi} P_{Xit} + b_{Yi} P_{Yit} + b_{Zi} P_{Zit}) \leq M_t, t = 1,2, \dots, T \quad (12)$$

$$\sum_{i=1}^n (b_{Xi} O_{Xit} + b_{Yi} O_{Yit} + b_{Zi} O_{Zit}) \leq \lambda_t^M M_t, t = 1,2, \dots, T \quad (13)$$

$$P_{Xit} + O_{Xit} \leq \Pi K_{Xit}, i = 1,2, \dots, n, t = 1,2, \dots, T \quad (14)$$

$$P_{Yit} + O_{Yit} \leq \Pi K_{Yit}, i = 1,2, \dots, n, t = 1,2, \dots, T \quad (15)$$

$$P_{Zit} + O_{Zit} \leq \Pi K_{Zit}, i = 1,2, \dots, n, t = 1,2, \dots, T \quad (16)$$

$$I_{ait}^s, I_{\beta it}^s, U_{it}^s, P_{Xit}, P_{Yit}, P_{Zit}, O_{Xit}, O_{Yit}, O_{Zit} \geq 0, i = 1,2, \dots, n, t = 1,2, \dots, T \quad (17)$$

where Π is a large positive number.

Constraint (7) determines either the quantity of finished products stored in the warehouse or the shortfall in meeting market demand. Constraint (8) determines the quantity of semi-finished products stored in the warehouse. The total quantity of semi-finished products produced at the company's plants during period t plus previous stock at period $t-1$ must equal the semi-finished products stored in the warehouse at period t plus the quantity of semi-finished products used to perform final assembly. The physical storage space at period t is limited by constraint (9). Constraints (10) and (11) limit the labor working hours during regular time and overtime respectively. Similarly, Constraints (12) and (13) limit the machining time during regular time and overtime respectively. Constraints (14) – (16) ensure that setup costs will be incurred when the corresponding production activities started. Constraint (17) ensures that the second-stage decision variables are non-negative.

3 Problem Solution

In order to illustrate the flexibility of the proposed stochastic programming approach for aggregate production planning problem for Christmas products, we use the data provided by the plush toy company in Hong Kong. The tactical/operational level of decision-making in the production planning process is described below. Based on the company's projection report, a two-month planning horizon is determined (November and December).

The company receives sales orders from its sales branches covering America and Europe. Each order may require two type of products, $i = 1,2$ covering 8 weeks, $t = 1,2,\dots,8$. It is assumed that future economic scenarios will fit into one of four possible scenarios – boom, good, fair and poor – with associated probabilities of 0.40, 0.25, 0.20 and 0.15 respectively.

The setup costs as well as labor and machine requirements of different production activities are given in Table 1. Table 2 shows the limitations on workforce level, machine capacity, overtime production and warehouse spaces in each period. Table 3 lists regular time labor cost, and hiring and laying-off costs associated with changes in the workforce level. The unit space occupation for finished products and semi-finished products are provided in Table 4. The production cost, inventory cost and shortage cost are shown in Tables 5–7. It is noted that, owing to the characteristics of the products, the shortage cost is time-sensitive and dramatically increases as the time approaches the event kick-off period (i.e. the ending period). For each weekly period, the product quantities required under different economic scenarios for the market are shown in Table 8.

The production loading plan with postponement strategy is shown in Table 9. It can be seen that the majority of products are produced using regular-time labor. In order to meet the growth of demand in the last two periods, production management is recommended to produce semi-finished products in periods 4–6 and perform final assembly in periods 7 and 8. The majority of resources consumed in period 8 are used to perform final assembly of product 2. It is shown that, using the postponement strategy, more products can be produced, particularly in period 8. Lastly, the workforce level in each period attains the upper-bound limit. The corresponding number of workers hired and laid off can also be found in Table 9.

The production loading plan without postponement strategy is also shown in Table 9. Compared with the production loading plan with postponement strategy, it is noted that the company produces more finished products and stores them in periods 5 and 6 for the demand in December. Since the storage of finished products incurs higher inventory cost and takes up more warehouse space, the production planning without postponement strategy is not a preferable for production management. Therefore, one of the advantages of postponement strategy is that, without adding extra costs and resources such as machine capacity, workforce level and warehouse space, more products can be produced in December with postponement strategy. This strategy is more attractive for production management.

The breakdown of costs incurred for production plans with and without postponement strategy is listed in Table 10. For the production planning with postponement strategy, the operational cost, which is the sum of production cost, setup cost, labor cost, inventory cost, and hiring cost and lay-off cost, is \$10,400,305. Clearly, when the demand requirements are smaller than the available production (from previous inventory and current production) the stock will be kept at the end of each particular period t under scenario s , and the corresponding inventory cost will be incurred. On the other hand, when the demand requirements are not satisfied, the company's service level and goodwill will be damaged. Compensation may be considered to cover the excess demand. This compensation is considered as a penalty cost. Table 10 shows that, under the optimal production

loading plan, the penalty cost is \$4,758,366. Overall, the total cost, which is the sum of operational cost and penalty cost, is \$15,158,671.

Originally, under the present strategy (without postponement) the total cost incurred is \$16,248,222. In comparison with the present strategy (without postponement) a saving of about 6.7% is made by following the proposed strategy (with postponement).

4 Conclusion

In this study, a stochastic programming approach for the aggregate production planning problem for perishable products with uncertain demand is proposed. The computation results obtained from a set of real-world data show that the proposed model is practical for dealing with uncertain economic scenarios. It is believed that the model can provide a credible and effective methodology for real-world production planning problems in an uncertain environment. However, there is still room for improvement and investigation. First, real data from companies can be used to validate the model and to analyze its sensitivity to changes in production planning strategies. Second, sensitivity analysis may be conducted on the cost parameters in the objective function to test the trade-off between total cost and shortage costs. Third, the selection of probability distribution of economic scenarios could be further investigated. Finally, the whole area of study associated with segregating market demand by region/country, and including different selling prices by region/country, can offer scope for making the APP a more useful basis for decision-making, in which we are not simply minimizing costs of production, etc., but are maximizing profit.

References:

- [1] Cheng T, Li J, Wan CLJ, Wang S. Postponement strategies in supply chain management. New York: Springer; 2010. doi: 10.1007/978- 1- 4419- 5837- 2
- [2] Weskamp, Christoph, Achim Koberstein, Frank Schwartz, Leena Suhl, and Stefan Voß, A two-stage stochastic programming approach for identifying optimal postponement strategies in supply chains with uncertain demand, *Omega*, Vol. 83, 2019, pp. 123-138.
- [3] G.B. Dantzig, Linear programming under uncertainty, *Management Science*, Vol.1, 1955, pp. 197–206.
- [4] P. Kall and S.W. Wallace, *Stochastic Programming*, John Wiley and Sons, 1994.
- [5] H.L. Lee and C. Billington, Design products and processes for postponement. In S. Dasu and C. Eastman(ed.) *Management of Design: Engineering and Management Perspectives*, Kluwer Publisher, 1994, pp. 105-122.
- [6] A. Ruszczyński and A. Shapiro, *Stochastic Programming (Handbooks in Operations Research and Management Science)*, Elsevier, 2003.
- [7] Jabbarzadeh, Armin, Michael Haughton, and Fahime Pourmehdi, A robust optimization model for efficient and green supply chain planning with postponement strategy, *International Journal of Production Economics*, Vol. 214, 2019, pp. 266-283.

Table 1. Operating and costs data (in HK\$, 1US\$ = 7.8HK\$).

	Product	Direct finished product production	Semi-finished product production	Transfer production
Setup cost (\$)	1	2000	1000	1500
	2	2500	1000	2000
Labor time (hour)	1	0.5	0.35	0.15
	2	0.6	0.35	0.25
Machining time (hour)	1	0.5	0.4	0.1
	2	0.6	0.4	0.2

Table 2. Warehouse, machine and workforce capacity.

Warehouse space limitation, \bar{I}_{tt}	1,000 m ³
Maximum workforce level, \bar{W}_t	1,000
Maximum machine capacity, M_t	16,000
Fraction of workforce available for over-time, λ_t^W ,	0.3
Fraction of machine capacity available for over-time, λ_t^M ,	0.4

Table 3. Labor costs and hiring and laying-off costs (in HK\$).

	Period							
	1	2	3	4	5	6	7	8
Labor cost per worker per period, C_t^W (\$)	80	80	80	80	80	80	80	80
Hiring cost per worker, C_t^H (\$)	80	80	100	100	100	80	80	80
Laying-off cost per worker, C_t^L (\$)	120	120	120	120	120	120	120	120

Table 4. Warehouse space occupation.

Product	Finished product warehouse space occupied (m ³)	Semi-finished product warehouse space occupied (m ³)
1	1.0	0.3
2	1.0	0.3

Table 5. Production costs under different scenarios (in HK\$).

Production cost		Product, <i>i</i>	Scenario, <i>s</i>			
			Boom	Good	Fair	Poor
Direct finished product production	Regular time	1	60	55	53	50
		2	70	65	63	60
	Overtime	1	60	55	53	50
		2	70	65	63	60
Semi-finished product production	Regular time	1	40	35	33	30
		2	40	35	33	30
	Overtime	1	40	35	33	30
		2	40	35	33	30
Transfer production	Regular time	1	40	35	33	30
		2	50	45	43	40
	Overtime	1	40	35	33	30
		2	50	45	43	40

Table 6. Inventory costs under different scenarios (in HK\$).

Inventory cost	Product, <i>i</i>	Scenario, <i>s</i>			
		Boom	Good	Fair	Poor
Direct finished product production	1	60	55	53	50
	2	60	55	53	50
Semi-finished product production	1	15	10	8	5
	2	15	10	8	5

Table 7. Shortage costs under different scenarios (in HK\$).

Product, <i>i</i>	Scenario, <i>s</i>	Period, <i>t</i>							
		1	2	3	4	5	6	7	8
1	Boom	400	440	484	532	584	644	708	780
	Good	300	330	363	399	438	483	531	585
	Fair	260	286	315	346	380	419	460	507
	Poor	200	220	242	266	292	322	354	390
2	Boom	480	528	580	640	716	772	852	936
	Good	360	396	435	480	537	579	639	702
	Fair	312	343	377	416	465	502	554	608
	Poor	240	264	290	320	358	386	426	468

Table 8. Market demand data.

Product, <i>i</i>	Scenario, <i>s</i>	Period, <i>t</i>							
		1	2	3	4	5	6	7	8
1	Boom	4000	4400	5000	5800	6800	8800	12600	23800
	Good	3000	3300	3750	4350	5100	6600	9450	17850
	Fair	2600	2860	3250	3770	4420	5720	8190	15470
	Poor	2000	2200	2500	2900	3400	4400	6300	11900
2	Boom	6400	6800	7600	8400	9600	11400	14200	20200
	Good	4800	5100	5700	6300	7200	8550	10650	15150
	Fair	4160	4420	4940	5460	6240	7140	9230	13130
	Poor	3200	3400	3800	4200	4800	5700	7100	10100

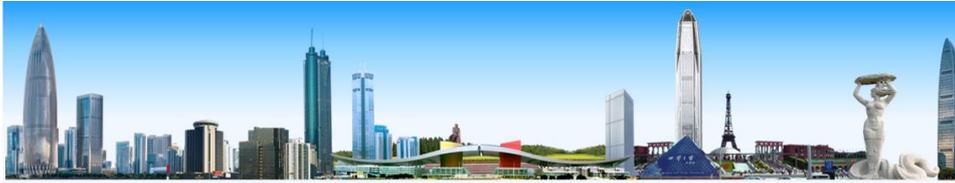
Table 9. Production loading plans with and without postponement strategy.

With postponement strategy										
		Product, <i>i</i>	Period, <i>t</i>							
			1	2	3	4	5	6	7	8
Direct finished product production	Regular time	1	2000	2860	1132	3770	4420	6600	7407	6443
		2	3200	2655	4940	5460	2240	7833	6650	0
	Overtime	1	0	0	2118	0	0	0	0	4800
		2	0	1765	0	0	4000	717	4000	0
Semi-finished product production	Regular time	1	0	0	0	0	8650	0	0	0
		2	0	0	0	804	4053	0	0	0
	Overtime	1	0	0	0	0	0	0	0	0
		2	0	0	0	4665	0	5629	0	0
Transfer production	Regular time	1	0	0	0	0	0	0	2043	6607
		2	0	0	0	0	0	0	0	15150
	Overtime	1	0	0	0	0	0	0	0	0
		2	0	0	0	0	0	0	0	0
Workforce level			441	441	441	680	1000	1000	1000	1000
Hiring			0	0	0	239	320	0	0	0
Laying-off			59	0	0	0	0	0	0	0

Without postponement strategy										
		Product, <i>i</i>	Period, <i>t</i>							
			1	2	3	4	5	6	7	8
Direct finished product production	Regular time	1	2000	2860	3250	3770	0	5720	8190	13050
		2	3200	3500	3175	3475	1333	8567	6508	5428
	Overtime	1	0	0	0	0	4420	0	0	4800
		2	0	920	1765	1985	317	4000	4000	0
Workforce level			441	441	441	496	1000	1000	1000	1000
Hiring			0	0	0	55	504	0	0	0
Laying-off			59	0	0	0	0	0	0	0

Table 10. Breakdown of costs (in HK\$).

	Production cost	Setup cost	Labor cost	Inventory cost	Hiring and laying-off	Operational cost	Penalty cost	Total cost
With postponement	6,695,583	42,500	480,323	3,118,974	62,925	10,400,305	4,758,366	15,158,671
Without postponement	6,177,306	36,000	465,600	4,018,201	62,925	10,760,032	5,488,190	16,248,222



Dynamic Optimal Approach for an Electric Taxi Fleet's Charging and Order-service Schemes

Kaize Yu, Pengyu Yan¹ and Zhibin Chen²

University of Electronic Science and Technology, Chengdu, China

NYU Shanghai, Shanghai, China

Corresponding author: Pengyu Yan, yanpy@uestc.edu.cn

Abstract: *This paper addresses an optimal charging and order service integration decision problem for an electric taxi (ET) fleet, operating by an e-platform. Compared with the conventional fuel taxis, the relatively long out-of-service time for recharging ETs significantly affects the revenue of drivers and the service level of the e-platform. In this study, we propose a charging-and-order-serve decision (COSD) system to jointly determine the ETs' dispatching and charging schemes. The dynamic optimization problem is formulated as a centralized multi-period stochastic model to maximize the total revenue of the fleet over a finite operational horizon. We develop an efficient algorithm based on the framework of rolling-horizon and considering uncertain orders of passengers. The result of the numerical experiment demonstrates the effectiveness of the proposed approach.*

Keywords: *Electric taxi; charging and order serving; multi-period stochastic model; dynamic decision; rolling horizon;*

1 Introduction

The electric vehicle (EV) has become an effective way to alleviate the environmental pollution caused by traditional fuel vehicles. Governments around the world are vigorously promoting electric vehicles. At the same time, a large number of electric vehicles are in operation on the online e-hailing platform. For example, an online hailing platform, Caocao has replaced all its vehicles with electric vehicles. Didi also plans to launch one million electric vehicles to provide travel services in 2020; Uber has proposed an Uber-green plan to replace traditional fuel vehicles with electric vehicles. In the following context, the vehicles operated on the platform are named as electric taxis (ETs for short). However, compared with fuel taxis, ETs have two main differences: First, ETs usually need to charge for one or two times during its operational time, due to the limited mileages in practice (200-300 km); Second, the charging time from empty to full power is about 0.5-5 hours depending on fast recharging modes, which implies that ETs have to quit from the e-platform (in off-line status) for relatively long time. The charging convenience of ETs is closely related to the number and layout of charging stations. In many cities, charging stations are insufficient comparing with the increasing number of ETs and the layout of these stations is not be well designed as well. Some stations in center or business areas are very busy and drivers have to wait in a long queue for charging their ETs, while in some remote areas, the utilizations of charging stations are usually low.

Let us consider a scenario that before the coming peak time, most ETs in a district (maybe the center of a city) may be in middle or low power. Independent drivers of these ETs decide to quit from the e-platform and go to nearby charging stations. In this situation, the service capacity (i.e., the number of available ETs) of the e-platform may decrease, which cannot satisfy the coming demands at peak time. Even worse, plenty of drivers going to the same stations may incur congestion in these stations. These drivers have to wait for a long time in the

queue, which further deteriorates the low service capacity of the e-platform at peak time. The charging operations of ETs may bring dramatic fluctuations of the service capacity of the e-platform. Therefore, the e-platform needs a centralized decision and scheme for ETs' service and charging operations considering the characteristics of ETs' charging operations and the current situation of charging stations. The illustration of the order-service and charging operations in an e-platform is presented in Figure 1.

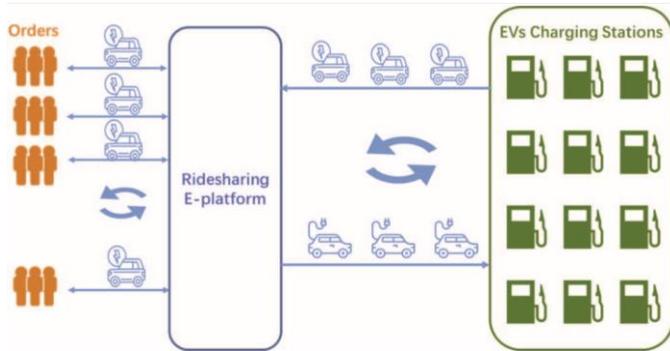


Figure 1: Illustration of order serve and charging operation in an e-hailing platform

2 Literature review

The charging problem of electric vehicles has been widely concerned in the transportation, operation optimization, and power engineering fields. However, most of the literature focuses on charging for private electric vehicles. There are few papers studied on the charging problem for an ETs fleet of the e-hailing platform. In the following, related research will be reviewed from two aspects: private electric vehicles charging problem and ETs charging problem.

2.1 Charging of private electric vehicles

At present, most of the research focuses on the charging problem of private electric vehicles and specifically considers the following two situations: charging in the community and charging in transit.

2.1.1 The problem of charging in the community

The community charging problem is usually considered to reduce the queuing and charging cost of electric vehicles under the constraints of grid resources. For a single charging station in a community, García-Álvarez, et al. (2018) and Hernández-Arauzo, et al. (2015) reduce the total charging delay under uncertain vehicle charging time and power supply capacity constraints. To coordinate charging station workload among different regions, Flath, et al. (2014) consider charging problem from the dimensions of time and space and stimulates drivers to select appropriate charging stations to balance the inter-regional grid load and charging waiting time based on price. Besides, some literature has studied how to coordinate the private electric vehicles to charge from the perspective of the power grid (Cao, et al., 2019; Umetani, et al., 2017; Wei, et al., 2014).

2.1.2 The problem of charging in transit

In addition to the research on charging in a community, some papers consider the charging in transit problem for private electric vehicles. Usually, in this kind of problem, the current electric power of electric vehicles is insufficient to support a journey, so it is necessary to select a charging station and charge for the electric vehicle on the way. To minimize the waiting time of electric vehicles, Qin, et al., (2011) have studied the selection of charging stations under a

given driving path. Under the same problem and objective function, Gusrialdi, et al., (2017) coordinates the selection of charging stations based on the collected traffic information and charging station status information. Sweda, et al., (2017) introduces charging cost into the selection of charging stations in a road network, that is, drivers reduce the charging cost caused by charging times and charging speed based on deciding where to charge and the amount of electricity.

Some papers (Schiffer, et al., 2017; Montoya, et al., 2017; Liao, et al., 2016) consider the VRP with charging station selection. He, et al., (2014) and Cen, et al., (2018) considered the problem of minimizing the charging cost under the condition of user equilibrium in the road network. In the case of vehicles charging in a road network, Sweda, et al., (2017) proposed an effective algorithm to find the optimal path and charging strategy based on the availability of power stations. Cao, et al., (2017) proposes a communication framework for electric vehicles, which recommends the relevant information of charging stations to drivers to help them choose the appropriate path and charging station. Baum, et al., (2019) takes into account a variety of charging station types, i.e. switching station, ordinary charging station and fast-charging station, as well as minimizing travel time with optional charging capacity.

2.2 The charging problem of e-hailing platform

The problem of e-hailing platform vehicles charging has not attracted enough attention, although e-hailing platforms are becoming more and more popular. There are obvious differences between e-hailing platform charging problem and private electric vehicle charging problem. First of all, e-hailing platform vehicles are serving passengers that charge is not allowed in the serving process. Secondly, the operation of an e-hailing platform is for profit, so it is necessary to consider not only how to reduce the cost factors such as charging costs and queuing time, but also serve more passengers to obtain more revenue. In this paper, we will review the related literature from two aspects: the optimization of e-hailing platform vehicle charging decisions and the joint optimization of order service and charging assignment.

2.2.1 Charging decision optimization

Some papers studied the model and algorithm to optimize charging costs for a single network. Tian, et al., (2016) predicts the current state of each ETs by detecting the historical charging data and real-time GPS trajectory and recommends the best charging station for ETs to minimize their driving distance and waiting time. Niu, et al., (2015) considered the overall charging coordination from the perspective of the ETs fleet, to minimize the total charging cost, charge station load, and maximize the utilization rate of charging equipment. Based on Niu's research, Yang, et al., (2015) considered the charging coordination optimization of ETs fleet from the perspective of the time-space dimension.

2.2.2 Joint optimize of charging and order serving

Yang, et al., (2018) proposed a two-stage charging coordination model. In the first stage, optimal charging time is selected based on the power state of electric vehicles, the time-varying income, and the queuing situation of charging stations; then, the appropriate charging stations are selected to reduce the queuing time through the game method. Ke, et al., (2019) considered a shared online platform where ETs and gasoline vehicles coexist and the market grows over time. ETs' drivers should manage their work and charging plan to balance the charging cost and operating income only from the time dimension. Sassi, et al., (2017) studied the problem of assigning orders for ETs and fuel vehicles when the passenger order information was completely confirmed. Besides, Hua, et al., (2019) considered an ET sharing platform to jointly

optimize the long-term charging station facility planning and real-time fleet operation (vehicle scheduling and charging decision-making) when the arrival of customers is uncertain.

Through the above review on private electric vehicles charging problem and e-hailing platform charging problem, we can find that most papers focus on the charging decision optimization of private electric vehicles in the community and transit. The optimization goal is always to minimize the charging cost of electric vehicles or the load of the community power grid. For the charging problem of the e-hailing platform, the existing parser only put forward the charging scheduling model and algorithm from the perspective of minimizing charging costs (including queuing time) of single ET. Some papers have studied the joint optimization of order service and charging assignment for the e-hailing platform but assuming that the order demand information is known. The optimization problem is established as a deterministic assignment scheduling model. Few papers can make decisions based on real-time information.

The remainder of the paper is organized as follows. Section 3 analysis the charging and order serving problem. And a multi-period stochastic model will be built. In Section 4, we convert the stochastic model into deterministic model based on the rolling-horizon framework. In Section 5, the performance of the model will discuss throughout the numerical experiment. And the conclusions are drawn in Section 6.

3 Problem analysis and formulation

In this section, we first analyze the proposed charging-and-order-serve decision (COSD) system, and then formally state the problem and corresponding assumptions.

3.1 Charging-and-order-serve decision system

The COSD system is a tactical level module embedded in the e-hailing platform, as illustrated in Figure 2. In this study, an e-hailing fleet with several electric taxis is considered. The real-time status of the fleet can be collected to e-platform which includes availability, location, state of charge (SOC), and so on. The status of ETs and charging stations can be monitored with the support of advanced technologies, such as IoT, GPS, etc. With collected real-time information, the COSD system will make the order-serving decision and charging decision jointly to maximize the total revenue of all ETs. More specifically, the system will determine which ETs should serve coming orders and which ETs should be charged. After the decision, the system will collect the information of arrival orders and assign ETs to serve them based on real-time matching and routing module and compute a price for passengers based on the pricing module. And ETs which should be charged will drive to a central charging station. The operation of the platform is complex and is the coordination of many modules. However, in this study, we are forced on the COSD system and simplify other modules.

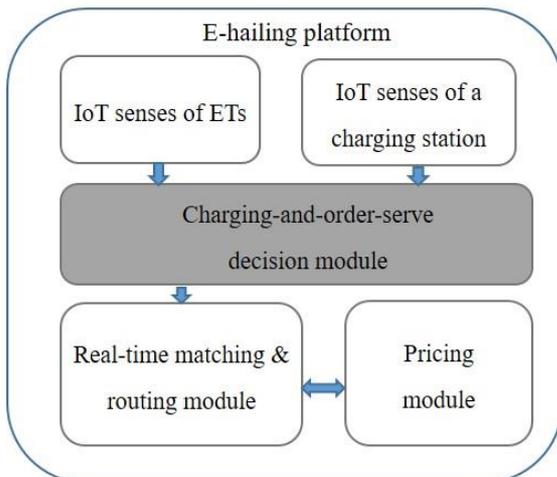


Figure 2: Illustration of an e-hailing platform with COSD system

3.2 Problem description

Let us consider an e-hailing platform that manages an electric taxi fleet to provide exclusive delivering service for passengers. The platform only operates in a finite time horizon and can be divided into T discrete periods. The period index is denoted as $t, t \in \{1, 2, \dots, T\}$. Under this discrete framework, at beginning of each period, the platform will make decisions based on the current state of all ETs and the workload of charging station. An ET can be one of the following statuses: 1) being available and wait to serve coming orders; 2) on served and 3) being charged at charging station. And we use A_t to represent all available ETs of fleet and P_t to represent all charging ETs at the beginning of period t . Normally, we have $|A_t| \cup |P_t| = E$, where E denotes the number of all ETs of fleet. Without loss of generality, we assume that ETs are homogeneous with the same battery capacity, and discretize into K levels. Through some IOT technologies, the platform can obtain the residual electricity of vehicles at any time. For an ET with $k, k \in \{1, \dots, K\}$ level, its SOC is in interval $\left(\frac{k-1}{K}, \frac{k}{K}\right]$ and an ET with 0 level means it cannot serve any passengers and need to be charged immediately. The district in which the fleet operates in is divided into I small cells that passengers always travel from one cell to another. The cell index is denoted as $i, i \in I$. We assume that there will be a charging station to provide exclusive charging serve for ETs in each cell. The ETs in the cell i only charge at the charging station of cell i . The time and SOC consumed from the location to the charging station can be ignored. Also, we assume that charging one SOC level will consume one period. Thus, the available ET set at period t is $S_t = \{s_{t,i,k} | 0 < k \leq K, i \in I\}$, where $a_{t,i,k}$ is the number of available ETs with level k in cell i . And the out-serve ETs is denoted as $P_t = \{p_{t,i,w} | 0 < w \leq K\}$, where $p_{t,i,w}$ is represent the number of ETs with type k in cell i at period t .

Similarly, passenger's orders can be classified into K types according to the requested power from the origin to the destination that passengers released. The order type index is denoted as l and computed by $l(i, j), i, j \in I$. In this study, the passengers are randomly arriving with unknown distribution. We use $F(\tilde{d}_{t,i,j}) = \{\tilde{d}_{t,i,j} | i, j \in I\}$ to represent the set of unknown probability distribution function, where $d_{t,i,j}$ represent the number of orders from cell i to cell j during period t , which are random variables. Besides, this study assumes that the order serve-time (i.e., ET traveling time) is an input parameter $\partial(i, j)$, depending on its origin and destination, which are estimated by the advanced real-time transportation information and navigation systems, such as Google map and Amap.

As aforementioned, the operational framework of COSD system is illustrated in Figure 3. At the beginning of period t , COSD system needs to make the charging decision $X_t = \{x_{t,i,k} | 0 \leq k < K, i \in I\}$ and the order-serve decision $Y_t = \{y_{t,i,j,l} | 0 \leq k \leq K, i, j \in I\}$. Specifically, $x_{t,i,k}$ represents the number of available ETs (i.e., S_t) in cell i with type k which need to be fully charged, and $y_{t,i,j,k}$ specifies the number of available ETs with type k to serve the special orders from cell i to cell j which are arrival during the current period. Note that an order with type $l(i, j)$ can only be served by type k ETs, which satisfy $k \geq l(i, j)$. During the period, some orders may be not served, thus we use $m_{t,i,j,k}$ to represent the number of orders travel cell i to cell j which are actually served by type k ETs. If the ETs driver served orders, they will obtain a reward $r(l(i, j))$. The reward has already been subtracted from the operation cost but not the charging fee. If there are no available ETs for serving order during period t , the order will lose and platform will suffer a penalty cost $\theta(l(i, j))$.

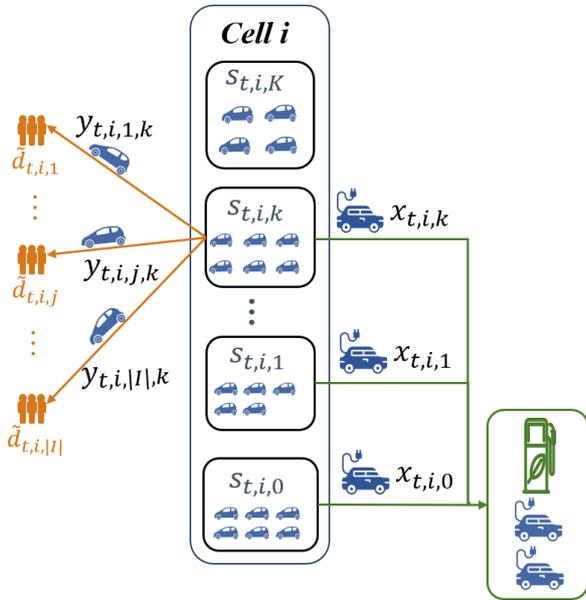


Figure 3: Operation framework of COSD system

To sum up, the problem addressed in this paper is to determine the ET charging and order serving decision in a finite horizon to maximize the total reward of the fleet. Notations frequently used throughout the paper are listed in Table 1. And the problem has the following important assumptions:

- An ET assigned to an order will pick up a passenger(s) immediately and travel to another cell. The SOC and time consuming from ETs location to passengers origin is ignored
- Each cell will have a charging station with N chargers. The travel time from the location of ETs to the station can be ignored. Furthermore, the station provides exclusively charging service for the ET fleet
- ET's drivers are employees of e-hailing platform and always comply with the instructions sent from the platform. In the numerical experiment section, a decentralized setting is addressed in which drivers are self-employees and want to maximize their revenues independently.
- Lastly, we consider impatient passengers. Passengers will switch to other transportation modes, such as subways, buses, or other e-hailing platforms if their orders are not confirmed in the current period.

Table 1: main notions in this paper

Indexes and parameters	
t	Period index, $1 \leq t \leq T$
k	ET type index, $0 \leq k \leq K$
I	Set of cells
S_t	Set of available ETs of period t
$s_{t,i,k}$	Number of available ETs of type k in cell i of period t
P_t	Set of charging ETs of period t
$p_{t,i,w}$	Number of ETs with type k in cell i of period t
$F(\tilde{d}_{t,i,j})$	Set of unknown probability distribution function of $\tilde{d}_{t,i,j}$
$\tilde{d}_{t,i,j}$	Number of random orders from cell i to cell j of period t , $i, j \in I$
$l(i, j)$	Type of orders from cell i to cell j
$r(l(i, j))$	Reward of serving an order with type $l(i, j)$
$\theta(l(i, j))$	Penalty of losing an order with type $l(i, j)$
$\partial(i, j)$	Consumed time from cell i to cell j
$m_{t,i,j,k}$	Number of type k ETs which are actually served order from cell i to cell j
Decision variable	
$x_{t,i,k}$	Number of ETs should be charged in cell i of period t
$y_{t,i,j,k}$	Number of ETs should serve coming order from cell i to cell j of period t

In this problem, ETs can be seen as a special kind of reusable resources. As illustrated in Figure 4, an ET with type k at period t will be unavailable until the beginning of period $t + K - k$, if it is assigned to be charged (dubbed as Case (a)). Or it will be available at the beginning of period $t + \partial(i, j)$, if it is assigned to serve order $l(i, j)$ (dubbed as Case (b)). However, different from the common reusable resource, such as equipment or machines, the ET's SOC changes when it is available once more, and its usability may thus be affected considerably. For example, the SOC is increased to K in Case (a), but it decreased to $k - l(i, j)$ in Case (b). On the other hand, as the orders randomly arrive during each period, the corresponding number of assigned ETs successfully serving orders (i.e., $m_{t,i,j,k}$) is thus a random variable affected by the random orders. In this paper, we are forced on the joint scheme of ETs charging and order serving. The order arrival sequence and the assigned rule are simplified.

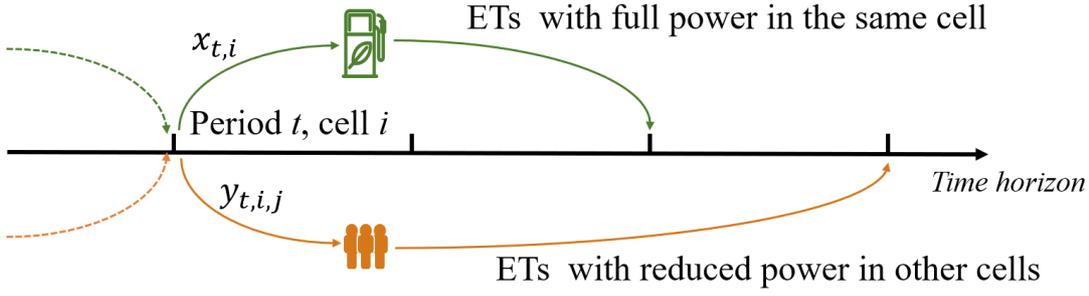


Figure 4: Illustration of ETs' reusable feature

3.3 Problem formulates

In this section, the problem will be formulated in a multi-period stochastic model.

3.3.1 Objective function

The objective of the model is to maximize the total reward of whole fleet over T horizon. Thus, the expected reward of period t is

$$r_t^s = \sum_{i \in I} \sum_{j \in I} \sum_{\tilde{d}_{t,i,j}=0}^{\infty} [r(l(i,j)) \times \sum_{k=l(i,j)}^K m_{t,i,j,k} - \theta(l(i,j)) \times (\tilde{d}_{t,i,j} - \sum_{k=l(i,j)}^K m_{t,i,j,k})^+] f_t(\tilde{d}_{t,i,j}) \quad (1)$$

where $r(l(i,j)) \times \sum_{k=l(i,j)}^K m_{t,i,j,k}$ are the reward of serving orders and $\theta(l(i,j)) \times (\tilde{d}_{t,i,j} - \sum_{k=l(i,j)}^K m_{t,i,j,k})^+$ is the penalty cost of the unsatisfied orders.

3.3.2 Constraints

Base on the problem statement in Section 3.2, four classes of constraints need to be considered in the problem.

(1) Capacity constraints of available ETs

Naturally, the ETs which assign to be charged and serve coming orders should less than the available ETs of different types, as shown in below constraints. Note that the ETs with full SOC don't need to charge and ETs with zero SOC cannot serve orders.

$$\sum_{l=1}^K y_{t,i,j,k} \leq s_{t,i,k}, \quad 1 \leq t \leq T, i \in I \quad (2)$$

$$x_{t,i,k} + \sum_{j \in I} y_{t,i,j,k} \leq s_{t,i,k}, \quad 1 \leq t \leq T, 1 \leq k < K, i \in I \quad (3)$$

$$x_{t,i,0} \leq s_{t,i,0}, \quad 1 \leq t \leq T, i \in I \quad (4)$$

(2) Served orders constraints

In our framework, the decision is made at the beginning of each period according to the expected order quantity before the order arrival. It likely happens that some passenger maybe not served if the real number of orders more than the planned ETs quantity, i.e., $d_{t,i,j} > \sum_{k=l(i,j)}^K y_{t,i,j,k}$, where $d_{t,i,j}$ is the actual order quantity from cell i to cell j . Thus, we have the following constraints:

$$\sum_{j \in I} m_{t,i,j,k} \leq \tilde{d}_{t,i,j}, \quad 1 \leq t \leq T, 1 \leq l(i,j) \leq K, i \in I \quad (5)$$

$$m_{t,i,j,k} \leq y_{t,i,j,k}, \quad 1 \leq t \leq T, 1 \leq l(i,j) \leq k \leq K, i \in I \quad (6)$$

(3) Capacity constraints of charging station

We assume that the charging station is in the center of the cell and have N chargers. Based on the IoT technology, the platform can monitor the workload of charging station. Thus, the ETS assign to be charged will less than the available chargers at period t .

$$\sum_{k=1}^{K-1} x_{t,i,k} \leq N - \sum_{w=1}^{K-1} p_{t,i,w}, \quad 1 \leq t \leq T, i \in I \quad (7)$$

(4) State transition equations

First, we consider the state transition of charging station. The number of ETs with type w in cell i at period $t + 1$ is the number of Ets with type $w - 1$ in cell i at period t add the number of ETs should be charged with type $w - 1$ in cell i at period t .

$$p_{t+1,i,w} = p_{t,i,w-1} + x_{t,i,w-1}, \quad 1 \leq w \leq K \quad (8)$$

For state transition of available ETs. The number of available ETs with type k in cell i at the beginning of period $t + 1$ is the number of available ETs with type k in cell i at period t minus ETs assign to serve ET with type k from cell i minus the number of ETs should be charged add number of ETs with type k from other cells. Also, for the number of type K ETs of the next period, it adds the number of ETs fully recharged at the charging station and becomes available again.

$$s_{t+1,i,K} = s_{t,i,K} - \sum_{j \in I} m_{t,i,j,K} + p_{t,K} \quad (9)$$

$$s_{t+1,i,k} = s_{t,i,k} - x_{t,i,k} - \sum_{j \in I} m_{t,i,j,k} + \sum_{j \in I} m_{(t-\partial(i,j)-1)^+, j, i, (l(i,j)+k)^+}, \quad 1 \leq k \leq K - 1 \quad (10)$$

$$s_{t+1,i,k} = s_{t,i,0} - x_{t,i,0} + \sum_{j \in I} m_{(t-\partial(i,j)-1)^+, j, i, l(i,j)^+} \quad (11)$$

To sum up, the multi-period stochastic model is formally presented as follows.

Problem P_S :

$$\max \mathbb{E}R^* = \sum_{t=1}^T r_t^s$$

s. t. (2) – (11)

4 Rolling-Horizon framework

In this section, a rolling-horizon framework will be applied to solve the multi-period stochastic problem. In actually, the probability distribution function of orders demand is difficult to estimate accurately. However, the historical data of orders can be easily obtained. Thus, we will first convert the stochastic model into a deterministic formulation based on the predicted means of order demand. And then, apply the rolling-horizon framework to solve the deterministic model with progressively issued orders.

4.1 Deterministic model

With historical data, the mean of orders can be easily computed and can be used to replace the random order demand variable. We use $\bar{d}_{t,i,j}$ to represent the mean of orders from cell i to cell j at period t and have the following expect reward and constrains:

$$r_t^d = \sum_{i \in I} \sum_{j \in I} [r(l(i,j)) \times \sum_{k=l(i,j)}^K m_{t,i,j,k} - \theta(l(i,j)) \times (\bar{d}_{t,i,j} - \sum_{k=l(i,j)}^K m_{t,i,j,k})^+] \quad (12)$$

$$\sum_{j \in I} m_{t,i,j,k} \leq \bar{d}_{t,i,j}, \quad 1 \leq t \leq T, \quad 1 \leq l(i,j) \leq K, \quad i \in I \quad (13)$$

Other constraints don't contain order random variables and are consistent with the corresponding constraints in the stochastic model. Thus, the deterministic model is built as :

Problem P_d :

$$\max \mathbb{E}R^* = \sum_{t=1}^T r_t^d$$

$$s. t. (2) - (4), (6) - (11), (13)$$

There are a variety of deterministic optimization techniques that can be directly used to solve the above problem P_d . In this paper, we adopt the state-of-art commercial optimization tool to solve the model. The solution quality and computational efficiency of the model are evaluated in Section 5.

4.2 Rolling-horizon algorithm

The rolling-horizon framework is widely used in multi-period decision problems and shows good performance. In this study, the order demands are released over the period. Once it is released, it will have no impact on subsequent decisions. The framework of rolling-horizon is shown in figure 5. The model will make the decision based on the orders mean of each following period. Before the next period, the number of arrival orders is released. The reward of current period can be computed by:

$$r_t = \sum_{i \in I} \sum_{j \in I} [r(l(i,j)) \times \sum_{k=l(i,j)}^K m_{t,i,j,k} - \theta(l(i,j)) \times (d_{t,i,j} - \sum_{k=l(i,j)}^K m_{t,i,j,k})^+]$$

Where $d_{t,i,j}$ is released the number of orders from cell i to cell j at period t . After that, the assignment decisions are implemented and the model will only be solved based on the order means of the following periods.

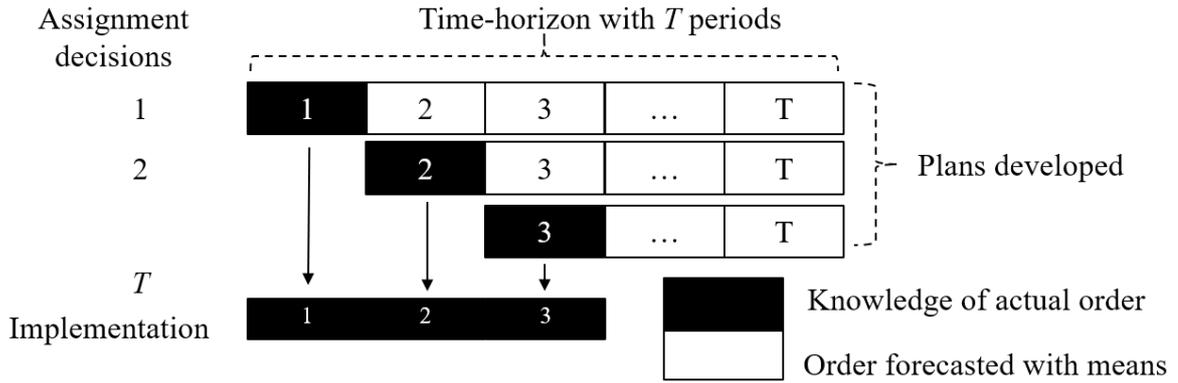


Figure 5: The rolling-horizon framework

To sum up, the algorithm of the rolling-horizon algorithm is shown below. At the beginning of the period t , the platform will collect the real-time information of available ETs $S_{t,i,k}$, $0 \leq k \leq K, i \in I$ and the workload of charging station $p_{t,i,k}$, $0 \leq k \leq K, i \in I$. Based on this information and order means $\bar{d}_{t',i,j}, t \leq t' \leq T$, the model will be solved. After that, the charging and order-serving decision of the current period will execute to obtain actually served order $m_{t,i,j,k}$ and revenue r_t based on the solution.

For period $t = 1$ to T

- Step 1: Platform collects real-time information: $S_{t,i,k}$ and $p_{t,i,k}$ at the beginning of period t ;
- Step 2: Solve model P_d with the mean $\bar{d}_{t',i,j}$ for coming periods $t', t \leq t' \leq T$;
- Step 3: Execute the charging and order-service assignments according to solution $x_{t,i,k}$, and $y_{t,i,j,k}$;
- Step 4: At the end of period t , the served demands $m_{t,i,j,k}$ are realized and the platform obtains the corresponding revenue r_t ;

End for

5 Numerical experiment

In this section, the performance of the model will be discussed. The historical order data were collected from Didi's GAIA open dataset in the center area of Chengdu City from November 1 to November 30, 2016. The region of order is divided into 9 cells, which is shown in figure 6. In the experiment, the operation horizon is from 6 a.m. to 12 p.m. the period length is set as 15 minutes. The parameters setting is shown in table 2. The deterministic model is solved by IBM ILOG CPLEX Optimization Studio V12.8.0 using a PC with i7 CPU @ 3.00GHz and 8.00G RAM.



Figure 6: Illustration of regional division and charging station location.

Table 2: numerical experiment parameters setting

$T = 108$ period
$K = 10$ level
$I = 9$ cells
$N = 20$ chargers
$E = 100$ ETs

First, we will discuss the performance of the model compared with the post-optimal solution and decentralized case. In the post-optimal solution, the model is solved based on all released order demands and we use $Gap2 = \frac{R^* - R^\wedge}{R^*}$ to represent the solution gap between our model and the post-optimal model, where R^* is the solution of post-optimal model and R^\wedge is the solution of our model. In the decentralized case, all drivers make charging decisions that are based on their knowledge and only go to enter the charging station where they are located. In this situation, it will happen that many ETs wait at the same charging station and waste their operation time. The arrival orders are randomly assigned the available ETs with the constrain $k \geq l(i, j)$. And the $Gap1 = \frac{R^* - R^{dec}}{R^*}$ to represent the solution gap between the decentralized solution and post-optimal solution, where R^{dec} is the solution to decentralized strategy. And the experiment simulates 20 weekdays based on Didi's historical data.

The result is shown in figure 7. It can be seen that our solution is far better than the solution of decentralized. The average gap of $Gap1$ is 232% and the average $Gap2$ is 47.1%.

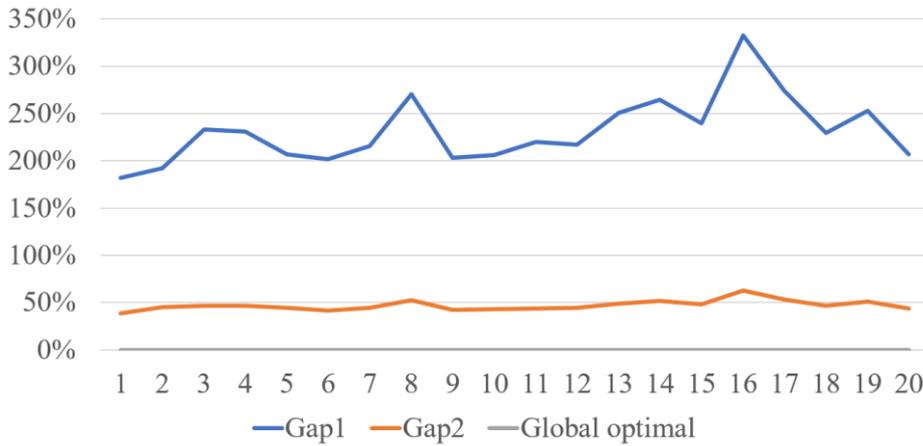


Figure 7: The result of Gap1 and Gap2

Then, we compare the Gap1 and Gap2 in different chargers. The result is shown in figure 8. It can be seen that our model is better than the decentralized decision in different chargers. And Gap1 and Gap2 will decrease with the number of chargers increase.

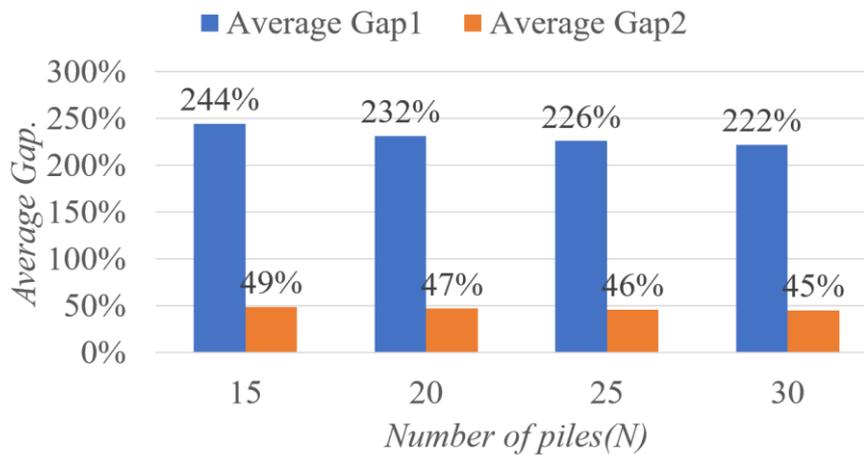


Figure 8: The changes of Gap1 and Gap2 under different chargers

6 Conclusion

In this paper, a charging and order serving problem is described. Compared with flue taxi, the milage of ETs are relatively short so that ETs' driver always recharges the battery during operation time which will waste their time, impact the platform serving capacity, and reduce their revenue. Thus, we propose a centralized COSD system to solve this problem. with the current information of available ET and charging station workload, the decision will be made based on a multi-period stochastic model. To solve the model, we convert it into a deterministic model under the rolling-horizon framework. During the numerical experiment, our model is far better than the decentralized strategy and shown a better performance with chargers increase.

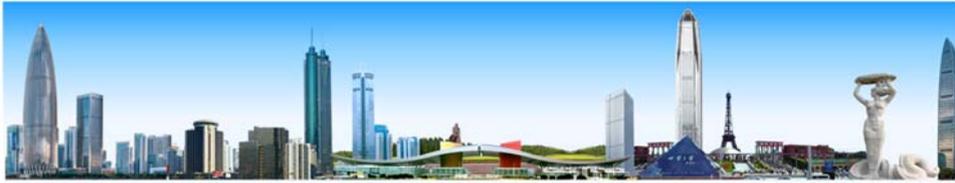
In future work, we will use the data-driven approach like SAA to replace the order means when converting the stochastic model into the deterministic. Also, the model can be built in the framework of the Markov Stochastic Process and use some state-of-art approaches like reinforcement learning to solve the problem. In this paper, we assume that drivers are unconditionally subject to decisions made by the platform. However, in practice, drivers have

their charging preferences. Therefore, in the following study, we will focus on how to set incentive prices to make drivers willing to obey the decisions made by the COSD system.

References

- García-Álvarez Jorge, González-Rodríguez Inés, Vela Camino R., et al. (2018): Genetic fuzzy schedules for charging electric vehicles, *Computers & Industrial Engineering*, 121(121): 51-61.
- Hernández-Arauzo Alejandro, Puente Jorge, Varela Ramiro, et al. (2015): Electric vehicle charging under power and balance constraints as dynamic scheduling, *Computers & Industrial Engineering*, 85(85): 306-315.
- Flath Christoph M., Ilg Jens P., Gottwalt Sebastian, et al.,(2014): Improving Electric Vehicle Charging Coordination Through Area Pricing, *Transportation Science*, 48(4): 619-634.
- Cao Yue, Kaiwartya Omprakash, Zhuang Yuan, et al., (2019) : A Decentralized Deadline-Driven Electric Vehicle Charging Recommendation, *Ieee Systems Journal*, 13(3): 3410-3421.
- Umetani Shunji, Fukushima Yuta, Morita Hiroshi, (2017): A linear programming based heuristic algorithm for charge and discharge scheduling of electric vehicles in a building energy Management system, *Omega-International Journal of Management Science*, 67(67): 115-122.
- Wei Lai, Guan Yongpei, (2014) : Optimal Control of Plug-In Hybrid Electric Vehicles with Market Impact and Risk Attitude, *Transportation Science*, 48(4): 467-482.
- Qin Hua, Zhang Wensheng, (2011) : Charging scheduling with minimal waiting in a network of electric vehicles and charging stations; proceedings of the Proceedings of the Eighth ACM international workshop on Vehicular inter-networking, F, ACM.
- Gusrialdi Azwirman, Qu Zhihua, Simaan Marwan A., (2017) : Distributed Scheduling and Cooperative Control for Charging of Electric Vehicles at Highway Service Stations, *IEEE Transactions on Intelligent Transportation Systems*, 18(10): 2713-2727.
- Swed Timothy M., Dolinskaya Irina S., Klabjan Diego, (2017) : Optimal Recharging Policies for Electric Vehicles, *Transportation Science*, 51(2): 457-479.
- Schiffer Maximilian, Walther Grit, (2017) : The electric location routing problem with time windows and partial recharging, *European Journal of Operational Research*, 260(3): 995-1013.
- Montoya Alejandro, Gueret Christelle, Mendoza Jorge E., et al. (2017) : The electric vehicle routing problem with nonlinear charging function, *Transportation Research Part B-Methodological*, 103(103): 87-110.
- Liao Chung-Shou, Lu Shang-Hung, Shen Zuo-Jun Max, (2016) : The electric vehicle touring problem, *Transportation Research Part B: Methodological*, 86(86): 163-180.
- He Fang, Yin Yafeng, Lawphongpanich Siriphong, (2014) : Network equilibrium models with battery electric vehicles, *Transportation Research Part B: Methodological*, 67(67): 306-319.
- Cen Xuekai, Lo Hong K., Li Lu, et al., (2018) : Modeling electric vehicles adoption for urban commute trips, *Transportation Research Part B-Methodological*, 117(117): 431-454.
- Sweda Timothy M, Dolinskaya Irina S, Klabjan Diego, (2017) : Adaptive routing and recharging policies for electric vehicles, *Transportation Science*, 51(4): 1326-1348.
- Cao Yue, Wang Ning, Kamel George, et al., (2017) : An Electric Vehicle Charging Management Scheme Based on Publish/Subscribe Communication Framework, *Ieee Systems Journal*, 11(3): 1822-1835.
- Baum Moritz, Dibbelt Julian, Gams Andreas, et al., (2019) : Shortest Feasible Paths with Charging Stops or Battery Electric Vehicles, *Transportation Science*, 53(6): 1627-1655
- Tian Zhiyong, Jung Taeho, Wang Yi, et al., (2016) : Real-time charging station recommendation system for electric-vehicle taxis, *IEEE Transactions on Intelligent Transportation Systems*, 17(11): 3098-3109.
- Niu Liyong, Zhang Di, (2015) : Charging guidance of electric taxis based on adaptive particle swarm optimization, *The Scientific World Journal*, 2015(2015): 9-19.

- Yang Yuqing, Zhang Weige, Niu Liyong, et al., (2015) : Coordinated charging strategy for electric taxis in temporal and spatial scale, *Energies*, 8(2): 1256-1272.
- Yang Zaiyue, Guo Tianci, You Pengcheng, et al., (2018) : Distributed Approach for Temporal-Spatial Charging Coordination of Plug-in Electric Taxi Fleet, *IEEE Transactions on Industrial Informatics*, 15(6): 3185-3195.
- Ke Jintao, Cen Xuekai, Yang Hai, et al. (2019) : Modelling drivers' working and recharging schedules in a ride-sourcing market with electric vehicles and gasoline vehicles, *Transportation Research Part E: Logistics and Transportation Review*, 125(125): 160-180.
- Sassi Ons, Oulamara Ammar, (2017) : Electric vehicle scheduling and optimal charging problem: complexity, exact and heuristic approaches, *International Journal of Production Research*, 55(2): 519-535.
- Hua Yikang, Zhao Dongfang, Wang Xin, et al.,(2019) : Joint infrastructure planning and fleet management for one-way electric car sharing under time-varying uncertain demand, *Transportation Research Part B: Methodological*,128(128): 185-206.



IPIC 2020 | 7th International Physical Internet Conference | Shenzhen

Analysis of a Physical Internet enabled parking slot management system

Bing Qing Tan¹, Suxiu Xu² and Kai Kang³

1. School of management, Jinan University, Guangzhou, China
2. Institute of Physical Internet, School of Intelligent Systems Science and Engineering, Jinan University (Zhuhai Campus), Zhuhai, China
3. Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Hong Kong

Corresponding author: tanbingqing0910@stu2018.jnu.edu.cn

Abstract: This paper considers a Physical Internet enabled parking slot management system where the drivers are matched with two types of parking slots through the Vickrey auction. In order to study the impact of the implementation of a system on the drivers and parking slots, a two-stage stochastic model is developed. Firstly, according to the number of parking slots and their cost function distributions, this paper calculates the expected price and the expected profit in the auction model. After that, a continuous-time Markov chain model is established to evaluate the performance in a dynamic environment, including random arrivals of the Poisson process and possible abandonment of drivers and parking slots. By combining these two models, this paper utilizes a quantitative method to analyze the performance of the system, and our analysis plays an important role in every part of the system in future decisions.

Keywords: Physical Internet; Markov chain; Vickrey auction; Performance evaluation

1 Introduction

Searching for an available parking space in many downtown areas has become a daily concern. The time spent on cruising often constitutes a substantial portion of travel time, because drivers usually keep on cycling the parking area until they found an empty parking space (Liu et al. (2014)). Both theoretical and empirical literature substantiates that constantly cruising commonly results in more traffic congestion and air pollution (Anderson and de Palma, 2004; Arnott and Inci, 2006; Glazer and Niskanen, 1992; Glazer et al., 1992; Shoup, 1997, 2006). For example, Shoup (2005) proposes that cruising may considerably inflate overall vehicle travel since it makes trips longer: some parking places in Los Angeles induce 3,000 extra vehicle kilometers per year for cruising, van Ommeren et al. (2012) utilize empirical evidence to investigate cruising for parking, and finds cruising for parking is more common with shopping and leisure than for work-related activities. To alleviate such traffic congestion and improve the convenience for drivers, some studies in parking issues have emphasized the reservation system of parking spaces (Hanif et al., 2010; Hashimoto et al., 2013; Kaspi et al., 2014; Liu et al., 2014).

The difficulty of searching for an available parking space is largely reduced through the smart parking management system. For example, Xiao et al. (2018b) propose a model-based practical framework to predict future occupancy from historical occupancy data alone. Wang and He (2011) develop a new prototype of a reservation-based smart parking system, and implement a parking reservation policy to balance the benefit of service providers and requirements from the users. In fact, most existing studies mainly focus on the architecture and framework of the parking management system. The management problem of parking space is complicated by

parking time required for different drivers. Hence, a key aspect of management should be considered is that each parking space can be split into several parking slots. Meanwhile, with the development of the sharing economy, some studies have considered how to realize private parking space sharing (Kong et al., 2018; Xiao et al., 2018a; Xu et al., 2016).

However, little attention has been devoted to the modeling and analysis of the system. There is enormous potential in evaluating and analyzing the performance of the system after modeling. In this paper, we consider combining the auction model with a continuous-time Markov model (CTMM) that capture important characteristics of the system. Also, we further evaluate the performance of the system for possible values of the whole parameters. This research is thus motivated by answering the following questions:

- (1) How to develop a simplified analytical model that captures important characteristics of the system?
- (2) The available time of parking spaces may not exactly match the needs of drivers. How to allocate and make price when space and drivers choose to abandon the platform?
- (3) How to consider performance evaluation from different aspects.

On this basis, we first develop a Physical Internet-enabled parking management system framework that integrates agent and IoT technologies. The system framework facilitates parking space trade between parking space owners and drivers, and also collects diverse data automatically for the effect analysis. Then, we consider there are two types of parking slots: public parking slots and private parking slots that have lower costs for the platform. Drivers prefer the private parking spaces. We also consider there are two types of drivers: who only need one parking slot and who need multiple consecutive parking slots. In order to model and analyze, we develop a general two-stochastic model of a parking slot management platform (PSMP). Firstly, according to the number of parking slots and their cost distributions, this paper calculates the expected price in the auction model. After that, a continuous-time Markov chain model is established to evaluate the performance in a dynamic environment, including random arrivals of the Poisson process and possible abandonment of drivers and parking slots. By combining these two models, this paper utilizes a quantitative method to analyze the performance of the system, and our analysis plays an important role in every part of the system in future decisions.

2 Related study

Parking plays a significant role in urban transport systems, many articles address one or both of modeling and evaluation for the parking problem. Most studies in the modeling category seek to shorten cruising time by giving guidance to motorists in parking systems. Auctions ask and answer the most fundamental questions in economics: who should get the goods and at what prices (Cramton et al., 2007). Hence, there has been a growing interest in using auctions for resource allocation and pricing problems. For instance, Edelman et al. (2007) investigated a new auction mechanism used by search engines to sell online advertising, and they proved this mechanism is an ex-post equilibrium, with the same payoff to all players as dominant strategy equilibrium of VCG mechanism. Two types of auction mechanisms were designed and compared for multi-unit transportation procurement, and allocated carriers to shippers efficiently in logistics e-marketplaces (Huang and Xu, 2013). A reverse iterative combinatorial auction was designed as the allocation mechanism to assign the spectrum resources for device-to-device communications with multiple user pairs (Xu et al., 2013). And in a literature review paper, Lafkihi et al. (2019) summarized a large number of studies that are related to transportation service procurement and further explained why auction is a suitable mechanism

for resource allocation and pricing problems. As a kind of necessary resource in our daily life, parking spaces are frequently utilized resources. Recently, more and more researchers applied auctions to efficiently allocate and reasonably make the price of parking spaces. Xiao et al. (2018a) addressed two truthful auction mechanisms to design the parking slot allocation and transaction payment rule based on full consideration of parking time assignment, and they compared two mechanisms to solve the four fundamental shared parking problems. They further propose a fair recurrent double VCG (FRD-VCG) auction mechanism to approach the emerging shared parking management problem, this mechanism has the potential to persuade participants to remain in the market whilst it improves the market's retention rate, the parking slot's utilization rate and the participants' utilities (Xiao and Xu, 2018). Kong et al. (2018) combined O-VCG auction with market design mechanisms, and they integrated this combination into parking space sharing and allocation problems. Tan et al. (2019) designed two sequential auction mechanisms based on first- and second-price auction and they utilized forecasted price to combine these two sequential auction mechanisms with two sharing rules to solve parking space allocation and pricing problem. And an auction-based implementation was proposed, to guarantee congestion-free traffic, and to reduce both the travel time and travel time unreliability. This auction-based highway reservation shows great potential as a new traffic management system (Su and Park, 2015). Shao et al. (2020) considered an auction-based parking reservation problem where a parking management platform can deal with the demand disturbances based on the proposed effective multi-stage Vickrey-Clarke-Groves (MS-VCG) auction mechanism.

3 Physical Internet-enabled urban parking management platform

Figure 1 describes the system architecture of the PI-enabled parking management system. As a classical parking management solution, this system consists of three levels: Infrastructure as a Service (IaaS) level, Platform as a Service (PaaS) level and Software as a Service (SaaS) level.

IaaS level contains hardware and software layers. The hardware layer includes smart sensors, smart gateway, servers and storage, and networks. PI-enabled smart parking environment is established using smart sensors, such as RFID for parking access control, closed-circuit television for parking surveillance and security, and ultrasonic sensors for parking vacancy detection. Physical parking objects, such as parking spaces and security barriers, are converted into smart parking objects. Smart gateway connects, manages and controls smart parking objects. Networks transmit parking-related data in a flexible-to-configure manner, and generally consist of 5G, WiFi, Bluetooth, Transmission Control Protocol/Internet Protocol, Ultra-wideband and Zigbee. In addition, the software layer includes a Gateway Operating System (GOS) and management tools. GOS as a light-weighted middleware system is deployed on desktops, servers and mobile devices to manage smart objects. Smart parking objects are thus managed in different parking scenarios, such as curbside parking areas, public and automated parking garages, and private parking spaces.

PaaS level contains five key components: agent parking space repository, platform service management, parking space agent management, data analytics services, and database. These components facilitate easy deployment and reduce the complexity of managing the underlying hardware and software. Relying on smart parking objects, parking space agents are virtualized and stored in the agent parking space repository. The real-time data are also stored in the database. Moreover, the platform service management module allows the system owners to maintain and configure the system and its services for fulfilling the requirements of stakeholders. The agents are managed by the parking space agent management module for daily parking operations. Data analytics services provide system owners with powerful tools to

measure the system performances. Based on the analysis results, the system owners could then propose some strategies and policies to improve the performance and increase the profit.

At the top of this architecture, SaaS level contains various applications for parking management, including parking supervision application, parking space allocation application, parking space pricing application, and parking space navigation application. Various stakeholders, such as drivers and parking space owners, access the applications through SaaS level. Other systems could also gain access to the system. For example, traffic control systems extract parking data to control traffic signals. Also, online payment applications are connected with the system to support online transactions. The applications not only improve driver satisfaction but also allocate parking spaces to drivers with reasonable prices, so that parking space owners make a healthy profit in the market.

In this paper, we aim at the performance analysis of the parking space market by data analytics services considering parking space allocation and pricing services. Furthermore, the PI-enabled parking management system collects required data for analysis, including the arrivals of both parking space owners and drivers, their abandonment and payment records.

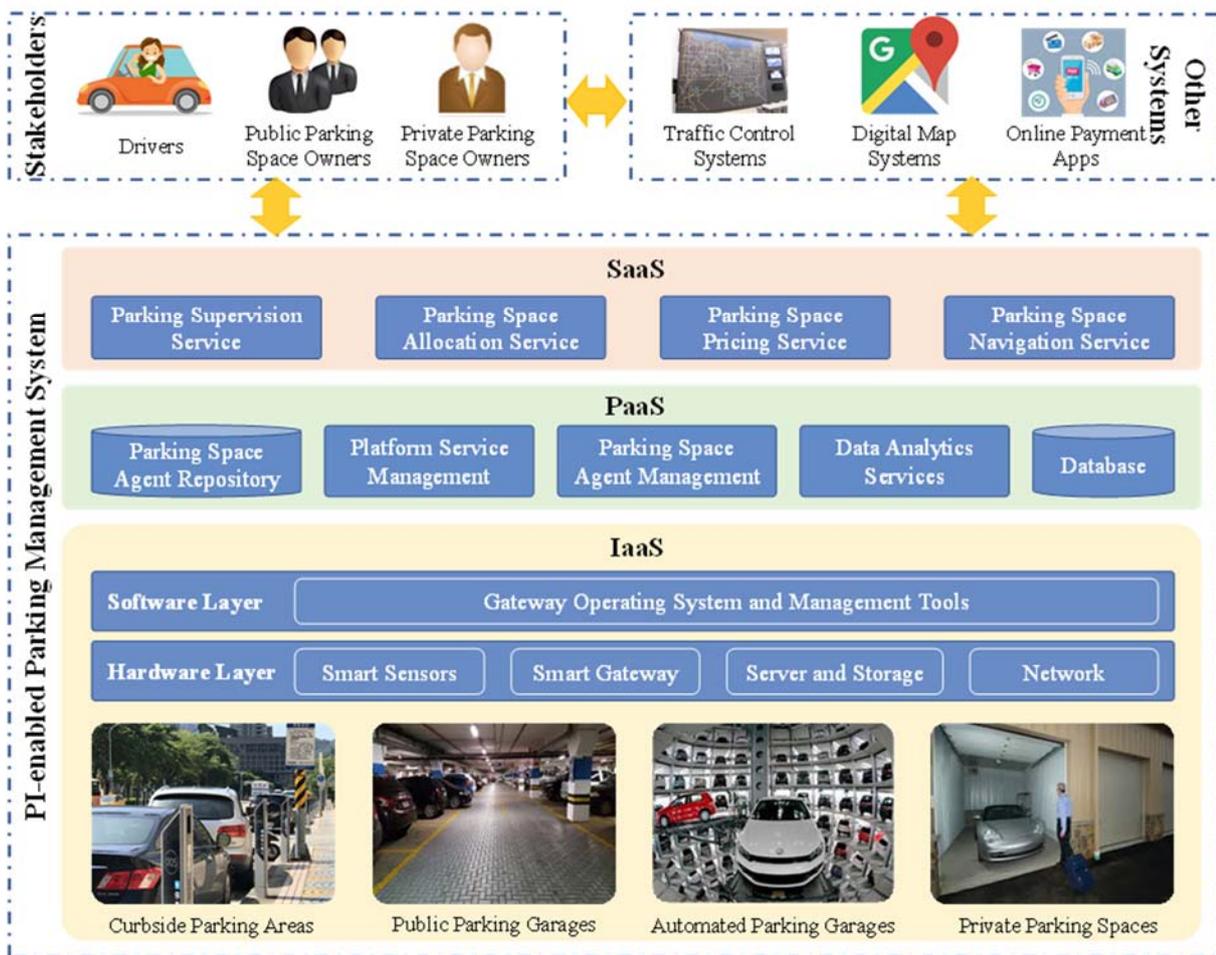


Figure 1: Architecture of the Physical Internet-enabled parking management systems

4 Model

4.1 Problem description

We consider a Physical Internet-enabled parking slot management platform (PSMP) where a number of available parking slots are allocated to drivers. In the platform, there are two types of parking slots that respond to two types of drivers by utilizing auctions to make short-term contracts. Two types of parking slots are public parking slots and private parking slots that have lower costs for the platform. Each parking space is split into several parking slots based on their available time, e.g. 8:00 am to 8:30 am, 4:00 pm to 4:30 pm. Two types of drivers are those who only need one parking slot and who need multiple consecutive parking slots. The parking orders from drivers are matched with parking slots through an electronic auction based on the auction centers (ACs) that are commonly shared among drivers and parking slots. The parking slots with the same available time and in the nearby area are regarded as homogeneous, so each auction center is designed to handle the matching of drivers and parking slots in a specific area. The parking demand is given to the parking space owner who submits the lowest bid. If no parking space is available or no appropriate bid is submitted, the driver will simply withdraw the parking demand from the platform. Also, parking spaces which stay at the platform may be useless after some time if they fail to be allocated to drivers.

4.2 Assumption

The following assumptions are made about the drivers and parking slots.

- (1) One type of drivers only need one parking slot i.e. $i = 1$, and another type of drivers require two consecutive parking slots i.e. $i = 2$.
- (2) Two types of drivers arrive randomly to the platform following a Poisson process with rate parameter λ_1, λ_2 , respectively.
- (3) If there are no parking slots are available, two types of drivers randomly abandon following a Poisson process with rate parameter α_1, α_2 , respectively.
- (4) There is no collusion between drivers.
- (5) Two types of parking spaces are public parking spaces and private parking spaces, and the cost of private parking spaces is lower than the cost of public parking spaces, i.e. $c_2 < c_1$ (Xu et al., 2016).
- (6) A parking space can be split into several parking slots.
- (7) Two types of parking slots arrive randomly to the platform following a Poisson process with rate parameter s_1, s_2 , respectively.
- (8) If there are no drivers are available, two types of parking slots are randomly failed to allocate following a Poisson process with rate parameter β_1, β_2 , respectively.
- (9) The per-unit cost c_j of two types of parking slots is drawn independently from a continuously differentiable distribution function $F_j(x)$ from a support $[c_j^-, c_j^+]$ with mean \tilde{c}_j , which is common to all the participants. In addition, we assume $c_2^+ < c_1^-$ based on the assumption (5).

4.3 Analysis of auction

Vickrey (Vickrey, 1961) discussed the bidding for a prize in which the highest bidder wins the prize but has to pay the second-highest price. Actually, this single-item second-price sealed-bid

auction is equivalent to Vickrey auction. In this paper, the driver with the lowest bid wins the parking slot and is paid the second-lowest price.

Denote $p(n_1, n_2, n_{s_1}, n_{s_2})$ as the expected price when n_i drivers and n_{s_j} parking slots are at the platform in steady-state. Let $q(n_1, n_2, n_{s_1}, n_{s_2})$ be the expected profit of the winner. In this paper, we consider the cost distribution of private and public parking slots are uniform. That is,

$$F_i(x) = \frac{x - c_i^-}{c_i^+ - c_i^-}, \quad c_i^- \leq x \leq c_i^+, \quad i = 1, 2.$$

$p(n_1, n_2, n_{s_1}, n_{s_2})$ and $q(n_1, n_2, n_{s_1}, n_{s_2})$ can be written as

$$p(n_1, n_2, n_{s_1}, n_{s_2}) = \begin{cases} c_2^- + \frac{2(c_2^+ - c_2^-)}{n_{s_2} + 1} & n_1 = n_2 = 0, n_{s_2} > 1 \\ c_1^- + \frac{c_1^+ - c_1^-}{n_{s_1} + 1} & n_1 = n_2 = 0, n_{s_1} > 1, n_{s_2} = 1 \\ c_1^- + \frac{2(c_1^+ - c_1^-)}{n_{s_1} + 1} & n_1 = n_2 = 0, n_{s_1} > 1, n_{s_2} = 0 \\ c_1^+ & n_1 = n_2 = 0, n_{s_1} = 0, n_{s_2} = 1 \\ & n_1 = n_2 = 0, n_{s_1} = 1, n_{s_2} = 0 \\ & n_1 + n_2 \geq 1, n_{s_1} = n_{s_2} = 0 \end{cases} \quad (1)$$

$$q(n_1, n_2, n_{s_1}, n_{s_2}) = \begin{cases} \frac{c_2^+ - c_2^-}{n_{s_2} + 1} & n_1 = n_2 = 0, n_{s_2} > 1 \\ c_1^- + \frac{c_1^+ - c_1^-}{n_{s_1} + 1} - \tilde{c}_2 & n_1 = n_2 = 0, n_{s_1} > 1, n_{s_2} = 1 \\ \frac{c_1^+ - c_1^-}{n_{s_1} + 1} & n_1 = n_2 = 0, n_{s_1} > 1, n_{s_2} = 0 \\ c_1^+ - \tilde{c}_2 & n_1 = n_2 = 0, n_{s_1} = 0, n_{s_2} = 1 \\ c_1^+ - \tilde{c}_1 & n_1 = n_2 = 0, n_{s_1} = 1, n_{s_2} = 0 \end{cases} \quad (2)$$

4.4 Analysis of the arrival-departure processes

The previous analysis has shown that the auction price and the profit depend on the number of drivers and the number of parking slots. Since the number of drivers and parking slots change dynamically with the random arrival-departure process of drivers and parking slots, we then examine the dynamics of the system to determine the steady-state distribution of the number of drivers and the number of parking slots at the platform.

Denote the state of the auction center at time t by $S(t) = (N_1(t), N_2(t), N_{s_1}(t), N_{s_2}(t))$. The process $\{S(t), t \geq 0\}$ is a continuous-time Markov Chain, because the inter-arrival and departure times of drivers and parking slots are exponentially distributed. This birth-death process is ergodic and has a stationary distribution. The steady-state probabilities are defined as

$$\pi(n_1, n_2, n_{s_1}, n_{s_2}) = \lim_{t \rightarrow \infty} \Pr\{N_1(t) = n_1, N_2(t) = n_2, N_{s_1}(t) = n_{s_1}, N_{s_2}(t) = n_{s_2}\} \quad (3)$$

In order to solve the model, the state-transition equations are obtained and the steady-state probabilities are calculated by truncating the state slot at state $(K_1, K_2, K_{s_1}, K_{s_2})$. K_1 is the maximum number of type 1 drivers that can be accepted by the platform, K_2 is the maximum number of type 2 drivers that can be accepted by the platform, K_{s_1} is the maximum number of type 1 parking slots arrive at the platform, and K_{s_2} is the maximum number of type 2 parking slots arrive at the platform. By setting sufficiently large values of K_1, K_2, K_{s_1} and K_{s_2} , the rejection probability for drivers and parking slots can be negligible. In addition, there are a total of $(K_1 + 1)(K_2 + 1) + (K_{s_1} + 1)(K_{s_2} + 1) - 1$ states in the resulting state $S(t) = (N_1(t), N_2(t), N_{s_1}(t), N_{s_2}(t))$.

5 Numerical study

5.1 Parameter setting

We assume $c_2^- = \tau c_1^-$, $c_2^+ = \tau c_1^+$ and $c_1^- = \varepsilon c_1^+$. Therefore, the distributions are defined in the range $[\varepsilon c_1^+, c_1^+]$ and $[\tau \varepsilon c_1^+, \tau c_1^+]$, the mean cost $\tilde{c}_1 = \frac{(1 + \varepsilon)c_1^-}{2}$ and $\tilde{c}_2 = \frac{(1 + \varepsilon)\tau c_1^+}{2}$. In order to ensure $c_2^+ < c_1^-$ holds, τ should be less than ε . And we set capacity for two types of drivers and two types of parking slots $K_1 = K_2 = 5$ and $K_{s_1} = K_{s_2} = 6$. We evaluate the performance of the system as indexes vary on the following two indexes for different $\lambda_1, \alpha_2, s_1, \beta_2, \tau$ and ε : (1) \bar{P} : the mean of expected price \bar{P} ; (2) \bar{Q} : the mean of expected profit.

5.2 Results

Drivers and parking spaces are traded at market price c_1^+ without PSMP. The auction center of PSMP is expected to lower the price paid by drivers. Table 1 shows the mean of the expected price paid by drivers for different type 1 driver and type 1 parking slots arrival rate, type 2 driver and type 2 parking slot abandonment rate, and the different cost distributions. Table 1 expresses that as the type 1 parking slots arrival rate s_1 increases, the number of parking slots participating in the auction will increase, which will lead to a decrease in the expected price. After the number of parking slots no longer increases, as the type 1 driver arrival rate λ_1 increases, the expected price will increase. Meanwhile, if the abandonment rate α_2 of type 2 driver increases, the total number of drivers remaining to continue bidding for parking spaces will decrease, which will result in a lower expected price. The larger the type 2 parking slot abandonment rate α_2 , the smaller the number of available parking slots in the system, and the result is a significant increase in the price. Especially when $\alpha_2 = 2$, the expected price exceeds 0.9 which is close to the market price $c_1^+ = 1$.

Table 1: The mean of the expected price

\bar{P}	λ_1				
s_1	0.2	0.3	0.4	0.5	0.6
1.2	0.21416	0.33214	0.45765	0.58254	0.69941
1.4	0.1876	0.29421	0.41	0.52838	0.64273
1.6	0.1675	0.26423	0.37062	0.48165	0.59168
1.8	0.15219	0.24061	0.33835	0.44183	0.54645
2	0.14038	0.22197	0.31207	0.40825	0.50693
\bar{P}	α_2				
β_2	0.1	0.2	0.3	0.4	0.5
0.4	0.48034	0.47221	0.46551	0.45997	0.45534
0.6	0.61336	0.60482	0.59767	0.59168	0.58663
0.8	0.72658	0.7189	0.71238	0.70686	0.70215
1	0.82367	0.81777	0.8127	0.80835	0.80459
1.2	0.9078	0.90431	0.90127	0.89862	0.89628
\bar{P}	ε				
τ	0.7	0.75	0.8	0.85	0.9
0.2	0.515	0.51097	0.50693	0.50328	0.4995
0.3	0.51742	0.51294	0.50847	0.50441	0.50022
0.4	0.51984	0.51492	0.51	0.50554	0.50093
0.5	0.52225	0.51689	0.51153	0.50668	0.50165
0.6	0.52467	0.51887	0.51307	0.50781	0.50237

Table 2 analyzes the effect of different type 1 driver and type 1 parking slots arrival rate, type 2 driver and type 2 parking slot abandonment rate, and the different cost distributions on the mean of the expected profit. Since the reverse auction is utilized, when type 1 driver arrival rate λ_1 increases, the demand increases. Therefore, when the number of parking slots participating in the auction remains unchanged, the larger λ_1 , the greater the mean of the expected profit \bar{Q} . Also, as the type 1 parking slots arrival rate s_1 increases and λ_1 remains unchanged, which will lead to a decrease in the expected profit. When the abandonment rate α_2 increases, there are

still type 1 parking slots in the system that can participate in the auction, since the change in α_2 has no obvious impact on the expected profit. In addition, when we set $\beta_2 = 2\beta_1$, as the two types parking slots' abandonment rates β_1 and β_2 both increase, the number of available parking slots in the system will decrease, which will lead to a decline in the expected profit.

Table 2: The mean of the expected profit

\bar{Q}	λ_1				
s_1	0.2	0.3	0.4	0.5	0.6
1.2	0.52482	0.53683	0.54965	0.56172	0.57235
1.4	0.49333	0.50807	0.52073	0.532	0.54188
1.6	0.46428	0.48249	0.49566	0.50642	0.5156
1.8	0.43782	0.4595	0.47357	0.48415	0.49274
2	0.41415	0.4388	0.45391	0.46453	0.47272
\bar{Q}	α_2				
β_2	0.1	0.2	0.3	0.4	0.5
0.4	0.52314	0.52395	0.52495	0.52609	0.52735
0.6	0.51445	0.51462	0.51502	0.5156	0.51633
0.8	0.50909	0.50878	0.50872	0.50888	0.50921
1	0.50568	0.50499	0.50457	0.50439	0.5044
1.2	0.50344	0.50245	0.50174	0.50128	0.50103
\bar{Q}	ε				
τ	0.7	0.75	0.8	0.85	0.9
0.2	0.5316	0.50216	0.47272	0.44604	0.41844
0.3	0.48527	0.45583	0.4264	0.39972	0.37212
0.4	0.43895	0.40951	0.38007	0.35339	0.32579
0.5	0.39262	0.36318	0.33374	0.30706	0.27947
0.6	0.34629	0.31686	0.28742	0.26074	0.23314

6 Conclusion

The contribution of this paper is to consider a Physical Internet-enabled parking slot management system where a number of available parking slots are allocated to drivers in the truthful reverse auction price. And the system performances are analyzed by a continuous-time

Markov chain. Our results show how the arrival rate of drivers and parking slots, and the abandonment rate of drivers and parking slots affect the expected price and the expected profit.

In addition, even though we focus on the parking slot management system, our model and results can be utilized in other situations such as online auction and transportation service procurement auctions where the number of bidders varies randomly. Moreover, combining auction and continuous-time Markov Chain can be applied to evaluate other resource management systems.

References

- Anderson, S.P., de Palma, A., (2004): The economics of pricing parking. *Journal of Urban Economics*, v55, no1, 1-20.
- Arnott, R., Inci, E., (2006): An integrated model of downtown parking and traffic congestion. *Journal of Urban Economics*, v60, no 3, 418-442.
- Cramton, P., Shoham, Y., Steinberg, R., (2007): An overview of combinatorial auctions. *ACM SIGecom Exchanges*, v7, no1, 3-14.
- Edelman, B., Ostrovsky, M., Schwarz, M., (2007): Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American economic review*, v97, no1, 242-259.
- Glazer, A., Niskanen, E., (1992): Parking fees and congestion. *Regional science and urban economics*, v22, no1, 123-132.
- Glazer, A., Niskanen, E., Economics, U., (1992): Parking fees and congestion. *Regional Science and Urban Economics*, v22, no1, 123-132.
- Hanif, N.H.H.M., Badiozaman, M.H., Daud, H., (2010): Smart parking reservation system using short message services (SMS), 2010 International Conference on Intelligent and Advanced System.
- Hashimoto, S., Kanamori, R., Ito, T., (2013): Auction-Based Parking Reservation System with Electricity Trading, 2013 IEEE 15th Conference on Business Informatics, pp. 33-40.
- Huang, G.Q., Xu, S.X., (2013): Truthful multi-unit transportation procurement auctions for logistics e-marketplaces. *Transportation Research Part B: Methodological*, v47, 127-148.
- Kaspi, M., Raviv, T., Tzur, M., (2014): Parking reservation policies in one-way vehicle sharing systems. *Transportation Research Part B: Methodological*, v62, 35-50.
- Kong, X.T.R., Xu, S.X., Cheng, M., Huang, G.Q., (2018): IoT-Enabled Parking Space Sharing and Allocation Mechanisms. *IEEE Transactions on Automation Science and Engineering*, v15, no 4, 1654-1664.
- Lafkihi, M., Pan, S., Ballot, E., (2019): Freight transportation service procurement: A literature review and future research opportunities in omnichannel E-commerce. *Transportation Research Part E-Logistics and Transportation Review*, v125, 348-365.
- Liu, W., Yang, H., Yin, Y., (2014): Expirable parking reservations for managing morning commute with parking space constraints. *Transportation Research Part C: Emerging Technologies*, v44, 185-201.
- Shao, S., Xu, S. X., Yang, H., Huang, G. Q., (2020): Parking reservation disturbances. *Transportation Research Part B: Methodological*, v135, 83-97.
- Shoup, D.C., 1997. The High Cost of Free Parking. *Journal of Planning Education and Research*, v17, 3-20.
- Shoup, D.C., (2006): Cruising for parking. *Transport Policy* 13, no6, 479-486.

- Su, P., Park, B.B., (2015): Auction-based highway reservation system an agent-based simulation study. *Transportation Research Part C: Emerging Technologies*, v60, 211-226.
- Tan, B.Q., Xu, S.X., Zhong, R., Cheng, M., Kang, K., (2019): Sequential auction based parking space sharing and pricing mechanism in the era of sharing economy. *Industrial Management & Data Systems*, v119, no8, 1734-1747.
- van Ommeren, J.N., Wentink, D., Rietveld, P., (2012): Empirical evidence on cruising for parking. *Transportation Research Part A: Policy and Practice*, v46, no1, 123-130.
- Wang, H., He, W., (2011): A Reservation-based Smart Parking System, 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPs). IEEE.
- Xiao, H., Xu, M., (2018): How to restrain participants opt out in shared parking market? A fair recurrent double auction approach. *Transportation Research Part C: Emerging Technologies*, v93, 36-61.
- Xiao, H., Xu, M., Gao, Z., (2018a): Shared parking problem: A novel truthful double auction mechanism approach. *Transportation Research Part B: Methodological*, v109, 40-69.
- Xiao, J., Lou, Y., Frisby, J., (2018b): How likely am I to find parking? – A practical model-based framework for predicting parking availability. *Transportation Research Part B: Methodological*, v112, 19-39.
- Xu, S.X., Cheng, M., Kong, X.T.R., Yang, H., Huang, G.Q., (2016): Private parking slot sharing. *Transportation Research Part B: Methodological*, v93, 596-617.
- Xu, C., Song, L., Han, Z., Zhao, Q., Wang, X., Cheng, X., Jiao, B., (2013): Efficiency Resource Allocation for Device-to-Device Underlay Communication Systems: A Reverse Iterative Combinatorial Auction Based Approach. *IEEE Journal on Selected Areas in Communications*, v31, no9, 348-358.
- Xu, S.X., Cheng, M., Kong, X.T.R., Yang, H., Huang, G.Q., (2016): Private parking slot sharing. *Transportation Research Part B: Methodological*, v93, 596-617.

Synchroperation in Industry 4.0 Manufacturing

Daqiang Guo^{1,2}, Mingxing Li¹, Zhongyuan Lyu¹, Kai Kang¹, Wei Wu¹, Ray Y. Zhong¹ and

George Q. Huang^{1*}

¹*Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Hong Kong, China*

²*Department of Mechanical and Energy Engineering, Southern University of Science and Technology, Shenzhen, China,*

Abstract

Industry 4.0 connotes a new industrial revolution with the convergence between physical and digital spaces, is revolutionizing the way that production operations are managed. The requirement of increased productivity, improved flexibility and resilience, and reduced cost in Industry 4.0 manufacturing calls for new paradigms that comply with the changing of production and operations management. In this paper, a concept of manufacturing synchroperation, refers to “*synchronized operations*” in an agile, resilient and cost-efficient way, with the spatiotemporal synchronization of men, machines and materials as well as data-driven decision-making, by creating, establishing and utilizing cyber-physical visibility and traceability in operations management, is proposed as a new paradigm of production and operations management for Industry 4.0 manufacturing. A Hyperconnected Physical Internet-enabled Smart Manufacturing Platform (HPISMP) is developed as a technical solution to support manufacturing synchroperation. Graduation intelligent Manufacturing System (GiMS) with “divide and conquer” principles is proposed to address the complex, stochastic, and dynamic nature of manufacturing for achieving synchroperation. An industrial case is carried out to validate the effectiveness of the proposed concept and method. This article provides insight into exploring production and operations management in the era of Industry 4.0.

Keywords: Industry 4.0; production and operations management; manufacturing synchroperation; Graduation Intelligent Manufacturing System (GiMS)

1. Introduction

Industry 4.0 connotes a new industrial revolution with the convergence between physical and digital spaces, which is triggered by the confluence of disruptive technologies, such as Internet of Things (IoT) (Xu et al., 2014), cyber-physical systems (CPS) (Lee et al., 2015), cloud computing (Xu 2012), big data (Kusiak 2017), digital twin (Tao et al., 2018), etc. With the support of these emerging technologies, traditional manufacturing resources have been converted into smart objects augmented with identification, sensing and network capabilities (Korteum et al., 2010). Thus, the dynamic production operations could be organized and managed in an integrated, optimized and synchronized manner with real-time information sharing and visibility (Guo et al., 2020a). The hyper-connection, digitization and sharing in the context of Industry 4.0 have the potential to revolutionize, or at least change, the way that production operations are done and therefore, how operations should be managed (Olsen and Tomlin, 2020).

The production and operations management has been shifted over the past fifty years, and three paradigms, including manufacturing collaboration, manufacturing interaction and manufacturing interoperation, can be classified with the enabling technologies and changing market. The paradigm of manufacturing collaboration aims at automating shop-floors by integrating different types of machines within a manufacturing company, which has rendered the feasibility of developing flexible manufacturing system (FMS) and computer-integrated manufacturing system (CIMS) (Buzacott and Yao, 1986; Mcgehee et al., 1994). Take the interaction of cross-organizational activities into account, the paradigm of manufacturing interaction extended production operations from a manufacturing company to a supply chain in a close-to-reality manner within a dynamic environment, which acts as key principles in agile manufacturing (AM) and networked manufacturing (NM) (Gunasekaran, 1999; Montreuil et al., 2000). More recently, to optimize the efficiency of production operations in the network, the paradigm of manufacturing interoperation is introduced, in which distributed manufacturing resources/capabilities can be interoperable in a close-loop network with timely production information exchange, which is the essence of cloud manufacturing (CM) and ubiquitous manufacturing (UM) (Xu 2012; Lin and Chen, 2017).

These paradigms for production and operations management are widely appreciated (Yin et al., 2018; Koh et al., 2019; Ivanov et al., 2020). The requirement of customized demand, increased productivity, improved flexibility and resilience, and reduced cost calls for more synchronized production and operations management that complies with changing business climate in Industry 4.0 manufacturing. Paving the way for transformation and implementation of Industry 4.0 manufacturing, major challenges still exist as follows.

(1) How to identify key characteristics for transformation and implementation of Industry 4.0 manufacturing, and derive a paradigm of production and operations management in the era of Industry 4.0 from these characteristics?

(2) How to leverage advanced technologies in the era of Industry 4.0 for developing effective architectures to support the transformation of the new production and operations management paradigm?

(3) How to cope with the complex, dynamic and stochastic nature of manufacturing by proposing effective methodologies to support the implementation of the new production and operations management paradigm?

The challenges mentioned above motivated this study and, therefore, the concept of manufacturing synchronoperation is proposed as a new paradigm of production and operations management in the era of Industry 4.0 with cyber-physical synchronization, data-driven decision synchronization and spatio-temporal synchronization. A Hyperconnected Physical Internet-enabled Smart Manufacturing Platform (HPISMP) assisted with digital twin and consortium blockchain, is developed as a technical solution to support the transformation of manufacturing synchronoperation. With the support of the HPISMP, Graduation Intelligent Manufacturing System (GiMS) with “divide and conquer” principles is proposed to address the complex, stochastic, and dynamic nature of manufacturing for achieving synchronoperation. An industrial case from an air conditioner manufacturer is carried out to illustrate the potential advantages of manufacturing synchronoperation.

The remainder of this paper is organized as follows. Related research streams are briefly reviewed in Section 2. In Section 3, the concept of manufacturing synchronoperation is introduced. A HPISMP is developed in Section 4. Section 5 presents GiMS with “divide and conquer” principles

for achieving synchronoperation. An industrial case from an air conditioner manufacturer is carried out in Section 6. Section 7 concludes the paper with some remarks on possible directions for future research.

2. Literature review

Many companies devote themselves to Industry 4.0 manufacturing. Siemens cloud-based IoT open operating system, MindSphere, connects products, plants, systems, and machines to enable industrial customers to harness the wealth of manufacturing data for decision-making (Siemens, 2020). GE IIoT platform, Predix, provides a complete solution for industrial data monitoring and event management, combining asset connectivity, and edge-to-cloud analytics processing to improve operational efficiency (GE, 2020). SAP cloud platform is designed to realize intelligent manufacturing that enables industrial customers to accelerate integration across the value chain while staying flexible and agile (SAP, 2020).

Industry 4.0 manufacturing with hyper-connection, digitization and sharing is revolutionizing production and operations management. This section briefly reviews the evolution of manufacturing paradigms based on enabling technologies and changing market at that time. Manufacturing paradigms can be classified into three types according to the principle of production and operations management. And challenges of these existing paradigms are then discussed and new requirements for transforming to Industry 4.0 manufacturing are proposed.

2.1. Manufacturing collaboration

Typical manufacturing paradigms has been developed to facilitate collaboration within a manufacturing company, including FMS and CIMS.

FMS refers to an integrated, computer-controlled complex of numerically controlled machine tools, automated material handling devices and computer hardware and software for the automatic random processing of palletized parts across various workstations (Buzacott and Yao, 1986). FMS utilizes the flexibility of job shops to simultaneously machine several part types to attain the efficiency of well-balanced, machine-paced transfer lines (Stecke, 1983). Eight types of flexibilities are summarized, including machine, process, product, routing, volume, expansion,

operation and production flexibility, to design an FMS (Browne et al., 1984).

CIMS refers to the harmonious connection and integration of automation equipment within a manufacturing facility (McGehee et al., 1994). CIMS utilizes computers and communication network to transform islands of enabling technologies into highly interconnected manufacturing system (Nagalingam and Lin, 1999) through a solution using STEP and STEP-NC for the integration of CAD, CAPP, CAM and CNC (Xu et al., 2005). CIMS improves data exchangeability and promotes adaptability of companies.

The key features in this era are: (1) They focus on a manufacturing company; (2) They aim at the automation of shop-floors by integration and collaboration of different types of machines; (3) They rarely integrate humans into manufacturing systems.

2.2. *Manufacturing interaction*

In the interaction era, typical paradigms are AM and NM. AM, originated from lean manufacturing, refers to the capability of surviving in a competitive environment of continuous and unpredictable change by reacting quickly to changing markets, driven by customer-designed products and services (Gunasekaran, 1999). The agility of manufacturing is provided through integrating reconfigurable resources with best practices in a knowledge-rich environment (Yusuf et al., 1999). A virtual enterprise is a typical application to characterize the global supply chain of a single product. It establishes the interaction with little liaison between companies to structure the whole system for agility (Martinez et al., 2001).

NM is the extension of agile manufacturing, and aims to collaboratively plan, control and manage daily activities and contingencies in a close-to-reality manner within a dynamic environment (Montreuil et al., 2000). NM increasingly focuses on information sharing that aims to create business relationships at different levels of shared information on price and capacity based on a distributed collaborative vision (D'Amours et al., 1999). Thus, standard information technology infrastructure is investigated to support cross-organizational activities for effective interaction (Akkermans and van der Horst, 2002).

The key characteristics in this era are: (1) The scope is extended from a company to a supply chain, and multiple companies are collaborated to manufacture a type of products; (2) Information

exchange plays a crucial role in the interaction between them; (3) Knowledge and wisdom of human are considered an important part of manufacturing systems.

2.3. *Manufacturing interoperation*

Recently, many manufacturing paradigms (e.g. IoT-enabled manufacturing, CM, and UM) have been designed to realize manufacturing interoperation using various technologies, such as IoT and cloud.

IoT-enabled manufacturing is an advanced principle where manufacturing resources are converted into smart ones able to sense, interconnect, and interact with each other to automatically and adaptively carry out manufacturing logics (Zhong et al., 2017). IoT provides manufacturing resources with the ability to exchange data and information real-timely (McFarlane et al., 2003). Open-loop networked manufacturing is thus closed, and tedious and error-prone manual data collection is eliminated, so that manufacturing resources can work together effectively (Huang et al., 2009). IoT lays the foundation of interoperation. Huang et al. (2008) utilize wireless manufacturing to manage work-in-progress inventories in job shops to retain existing operational flexibility while improving efficiency and capacity. Zhong et al. (2013) present a RFID-enabled manufacturing execution system to track and trace manufacturing resources and collect real-time data for making planning and scheduling decisions.

CM uses the network, cloud computing, service computing and manufacturing enabling technologies to transform manufacturing resources into services that are managed and operated in a unified way to share and circulate manufacturing resources (Zhang et al., 2014). Interoperability is the prerequisite for CM, because manufacturing resources need to be described, virtualized and integrated in a manufacturing cloud before sharing (Wang and Xu, 2013). Chen and Chiu (2017) find the smooth operation on cloud is hampered by interoperability when different cloud services are utilized. Wang et al. (2018b) classify interoperability in CM into four levels: data level, computing service level, manufacturing process level and CM service level. At manufacturing process level, a costing-based, generic deployment model is designed to identify the key process parameters that influence the interoperability of CM (Mourad et al., 2020). Also, synchronization as the extension of interoperation is proposed to address dynamics in production logistics

activities (Qu et al., 2016).

UM enables on-demand network access to a shared pool of configurable manufacturing resources but emphasizes the mobility and dispersion (Lin and Chen, 2017). Interoperability is also one of core characteristics, which closes the loop of production planning and control for adaptive decision-making (Zhang et al., 2011). Thus, manufacturing knowledge representation and data structure in UM need to be standardized, so that diversified UM systems are interoperable (Wang et al., 2018a). Wang et al. (2017) propose a function block-based integration mechanism to integrate various types of manufacturing facilities for interoperability in UM. Luo et al. (2017) present the synchronized production and logistics via ubiquitous computing to make real-time decisions within a factory.

The key characteristics in this era are: (1) Manufacturing resources worldwide are interoperable (Newman et al., 2008); (2) IoT facilitates information and data exchange to close the open-loop production process; (3) synchronization as an extension of interoperation is explored relying on real-time data to handle dynamics.

2.4. *Challenges*

From the literature, factors related to manufacturing paradigms in three eras are summarized in Table 1. Several key challenges are thus proposed for Industry 4.0 manufacturing. The first is what the general principle is and what its key characteristics are. Although Industry 4.0 has been applied to different application scenarios with various objectives, the essence of Industry 4.0 is rarely considered. The second is how state-of-the-art technologies can be fused and integrated to provide a technical solution for Industry 4.0 manufacturing. Many advanced technologies are proposed to support Industry 4.0 manufacturing, but the fusion of them is scarcely reported. The third is what approaches can be leveraged to address the complex, dynamic and stochastic nature of manufacturing optimization problems. As manufacturing systems become increasingly complex and stochastic, traditional methods can hardly provide optimal solutions timely in the context of Industry 4.0. Thus, this paper proposes the concept of synchroperation, introduces enabling technologies, and designs a methodology for synchroperation.

Table 1. Summary of manufacturing paradigms

Category	Manufacturing collaboration	Manufacturing interaction	Manufacturing interoperation
Production mode	Flexible production	Mass customization	Mass customization
Society needs	Variety of products	Customized products	Customized products
Market	Demand>Supply	Supply>Demand	Supply>Demand
Product volume	Small volume demand	Smaller volume demand	Fluctuating demand
Scope	Machine- machine level	Factory-factory level	Supply chain level
Business model	Push	Pull-Push	Pull
Technology enabler	Computer	Information technology	IoT, cloud computing

3. Manufacturing Synchroperation

3.1. Concept of manufacturing synchroperation

The requirement of customized demand, increased productivity, improved flexibility and resilience, and reduced cost calls for efficient production and operations management that complies with changing business climate in Industry 4.0 manufacturing. On the basis of the evolution of production and operations management paradigms, and from a manufacturing point of view, we understand synchroperation as a new paradigm of production and operations management in the era of Industry 4.0 as follows.

Synchroperation refers to “synchronized operations” in an agile, resilient and cost-efficient way, with the spatiotemporal synchronization of men, machines and materials as well as data-driven decision-making, by creating, establishing and utilizing cyber-physical visibility and traceability in operations management.

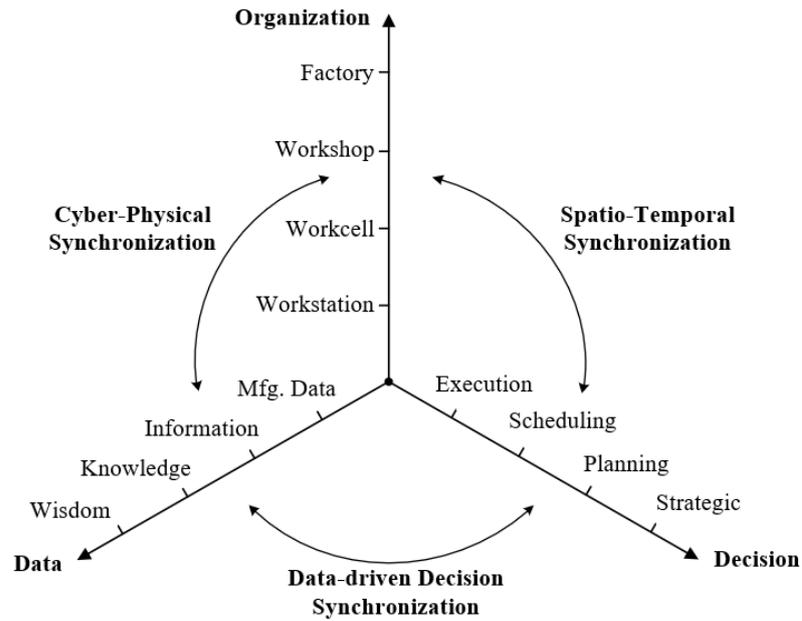


Fig.1. Overall framework for manufacturing synchroperation

Fig.1 shows the overall framework of manufacturing synchroperation with three key characteristics, including cyber-physical synchronization, spatio-temporal synchronization and data-driven decision synchronization. Cyber-physical synchronization focuses on the synchronization between cyber space and physical space through information visibility and traceability. The IoT and digital twin enabled ubiquitous connection, digitization and information-sharing in the context of Industry 4.0, present an opportunity for creating a digital equivalent representation of the physical entity (e.g., from small as a workstation, a workcell, to big as a workshop, a factory) and synchronizing them between cyber space and physical space with real-time information sharing.

Data-driven decision synchronization focuses on coordinated and global optimal production decisions benefiting from information sharing and data analytics. With enormous manufacturing data collected and shared in the cyber-physical system, valuable information and knowledge could be derived from the hidden patterns and correlations based on data mining, big data analytics and AI technologies. Thus, the coordinated, global optimal and even autonomous decision-making can be made for both short-term (e.g., scheduling and execution) and long-term (e.g., strategic and planning) production strategies.

For achieving successful implementation of these production strategies derived from

data-driven decision models, spatio-temporal synchronization is crucial as it focuses on decomposing complex production environment and operations into spatio-temporal units and synchronizing them in a “divide and conquer” manner. It not only ensures that the required production resources (e.g., men, machines and materials) could be allocated and utilized in the right place at the right time with synchronization of production operations, but also in turns decouples the decision models towards practical industrial application by significantly reducing the uncertainty, randomness and complexity.

3.2. *Synchroperability measures*

Following the concept of synchroperation, we define synchroperability as the ability of a manufacturing system to achieve synchronized operations. Synchroperability is a measure for synchroperation in manufacturing. Three important aspects of measures for synchroperation, including simultaneity, punctuality and cost-efficiency are derived from the literature, and will be considered comprehensively in this section.

Table 2. Measures for synchroperability

Measures	References	Environment	Major aims	Indicators
Simultaneity	(Hsu and Liu, 2009)	Job shop	Reduce finished product inventory level	Flow time
	(Luo et al., 2017)	Flow shop	Improve overall performance of production and logistics	
	(Chen et al., 2019)	Flow shop	Improve production efficiency	Waiting time
	(Guo et al., 2020b)	Job shop	Improve the synchronization degree between manufacturing and logistics	
Punctuality	(Chen et al., 2015)	Flow shop	Improve production lead time and shipment punctuality	Earliness and tardiness
	(Fazlollahtabar et al., 2015)	Job shop	Improve the performance of material handling system in production	
Cost-efficiency	(Qu et al., 2016)	Flow shop	Improve production-logistics resources utilization	Utilization
	(Lin et al., 2018)	Flow shop	Improve production efficiency	Setup time
	(Luo et al., 2019)	Flow shop	Improve overall performance of production and warehousing	Makespan
	(Lin et al., 2019)	Flow shop	Improve production efficiency	

As listed in Table 2, synchroperability measures are divided into three categories: simultaneity,

punctuality and cost-efficiency. Simultaneity is one of the most important aspects of measures for synchronoperation, and indicators for simultaneity, such as flow time and waiting time have been adopted in specific applications (Hsu and Liu, 2009; Luo et al., 2017; Chen et al., 2019; Guo et al., 2020b). Simultaneity concerns variation in completion times of jobs within the same package or order, which can be used to reduce finished product inventory level as well as improve production efficiency. Simultaneity could be considered as a measure in the production system that sensitive to job flow time or waiting time. For example, for producing large-size or fragile products, it is sensitive to job waiting time as holding such a product is quite expensive, which requires a measure of simultaneity in this production system (Guo et al., 2020c). Punctuality is another important aspect of measure that focuses on earliness and tardiness (Chen et al., 2015; Fazlollahtabar et al., 2015). Punctuality could be considered as a measure in the production system that advocates JIT production, as it can reduce the production lead time, inventory level and shipment punctuality (Chen et al., 2015). Cost-efficiency is a common aspect of measure for synchronoperation, most of the literature deals with such regular indicators as makespan, utilization and setup time (Qu et al., 2016; Lin et al., 2018; Luo et al., 2019; Lin et al., 2019). Cost-efficiency could be used as a measure in a complicated production environment that involves across multi-echelon and inter-organizational production activities. For example, for achieving overall optimization of make-to-order production and cross-docking warehouse, Luo et al (2019) proposed a synchronized production and warehouse decision model to minimize the overall makespan.

4. Synchronoperation Platform for cyber-physical traceability and visibility

To achieve the cyber-physical visibility and traceability serving for synchronoperation, a HPISMP, leveraging various IoT technologies, digital twins, big data techniques and consortium blockchain, is proposed in this paper. The overview of the technical framework is shown in Fig. 2.

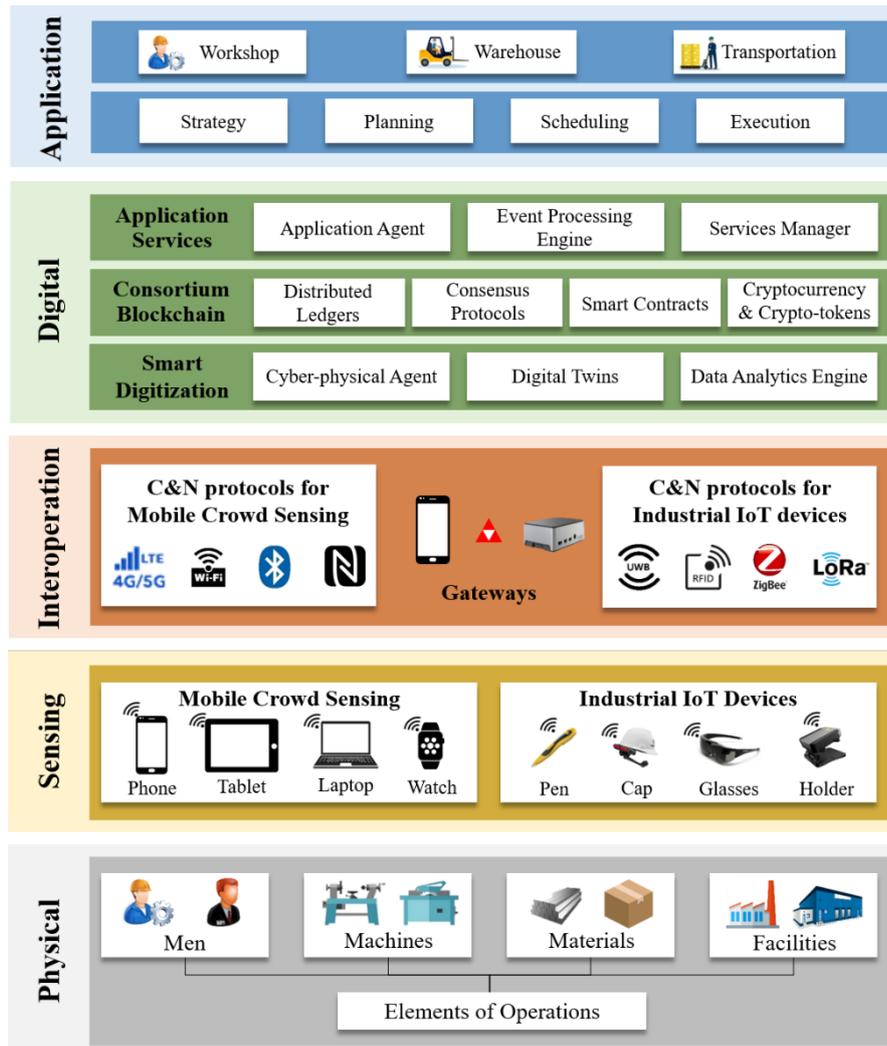


Fig. 2. Overview of hyperconnected Physical Internet-enabled smart manufacturing platform

The platform is divided into five layers, from bottom to top, namely physical, sensing, interoperation, digital, and application layer. Each layer is designed to connect, interact, and interoperate with each other so as to reinforce the overall synchronizability for production.

The first physical layer concerning men (e.g., managers and onsite operators), machines (e.g., production machines, vehicles and tools), materials (e.g., raw materials, Work-In-Processes (WIPs) and finished products), and facilities (e.g., production workshops and warehouses) that are fundamental elements of operations in manufacturing. Each type of element owns several categories classified by roles, functions, or phases. In line with actual demand, those elements may be equipped with electronic tags that contain a trickle of information primarily for identification in the cyber space, which can be captured passively or broadcast proactively (Zhao et al., 2017). Under this condition, it endows each element with capability of communication and interaction.

The second sensing layer includes a variety of IoT equipment that is used to enable physical objects with sensible, interactive, and intelligently reasoning capabilities. To be specific, mobile crowdsensing (MCS) is to collectively gather sensing data from nearby sensors via ubiquitous mobile devices, such as smartphones, tablets, laptops, and smartwatches (Ganti et al., 2011). Besides the application of MCS, Industrial Internet of Things (IIoT) devices are also widely deployed for the real-time data collection and transmission (Kong et al., 2020). Notably, wearable devices for men and machines enjoy high favour in the industry attributing to hands-free carry and convenient handling. For example, machines furnished with tag readers, like smart holders, are capable of detecting adjacent objects and triggering relevant events to facilitate operations. Else, men tend to carry smart devices to connect with peripheral objects and maintain responsive communication, such as versatile smart pens, caps, and glasses. Data or extracted information secured in this layer will be uploaded to the next layer for further processing.

The third interoperation layer aims to synchronize cyber-physical spaces and realize timely and seamless dual-way connectivity and interoperability between manufacturing objects and different application systems. It encompasses gateways and wireless communication and networking (WC&N) protocols. WC&N protocols serve as a data carrier to link with smart objects and the digital world (Zhang et al., 2011). In this platform, diversified wireless communication technologies are devised to be applied, such as 4/5G, LTE, Wi-Fi, Bluetooth, and NFC, which are typically provided in smartphones and personal wearables, and others like UWB, RFID, ZigBee, and LoRa that are preferably harnessed in the industry. All those protocols are embedded in gateways. Besides acting as a hardware hub, the gateway also offers a suite of software services, named gateway operating system, including definition, configuration, execution, and monitoring (Fang et al., 2013). Concretely, it can define the flow of data collection from heterogeneous devices and the cloud, configure essential setups and environmental conditions, execute data processing, information aggregation and exchanging, and, finally, monitor the entire operations for the malfunction detection. In addition, gateways have two types. One is the stationary gateways that are mounted at appointed spots and work in a plug-and-play way to ease the deployment. Another is the mobile gateways that are moveable and even portable since ubiquitous devices, like smartphones,

can install dedicated applications to serve as a gateway, which significantly extends the channel of data collection and reduces development cost.

The fourth digital layer is the cyberspace, in the form of the cloud or servers in reality, that replicates the physical world and adopts services-oriented governance. The main function of smart digitization (Lin et al., 2019) converts physical objects into cloud assets (Xu et al., 2018) and applies data analytics for different purposes. Thereinto, the cyber-physical agent acts as a gate of cyberspace to receive data from gateways. Based on those informative data, physical objects are mapped to digital twins under predefined logics. Else, the data refinement or pre-processing is implemented in the data analytics engine. The second module consortium Blockchain (Li et al., 2017) is used as a backbone to build up a database, facilitate workflow management, and ensure traceability, accountability, and transparency. In detail, the distributed ledgers serve as distributed databases where digital twins are stored and transactions recorded, and consensus protocols is a globally-agreed rule for distributed computing to control the access to the database and make sure ledgers trackable and irreversible. Any kind of codified procedures that refer to a series of real-life workflows are encrypted and stated in smart contracts to trigger events in order. Cryptocurrency or crypto-tokens working as an incentive mechanism is granted proportionally according to the quality and punctuality of task completion. The goal of the third module application services is to host and manage services committed to users. The application agent acts as another gate of the digital world to interact with user applications. Calling functions or taking responses pursuant to manipulations in the application behaves in the event processing engine. The services manager is devoted to administrating fundamental components of services and configuring logics between services.

The fifth application layer provides decision support systems and visualization tools to help conduct operations for participating stakeholders, including workshops, warehouses, and transportation in manufacturing. Each stakeholder is principally concerned with four kinds of applications, namely strategy, planning, scheduling, and execution. Additionally, applications are developed in forms of desktop and mobile terminal so that office staff can make decisions in front of the desktop, and operators simply bring mobile devices to accomplish tasks. Furthermore, those applications allow individuals from different departments to manipulate separately without

interference but for the same goal among the whole factory in real time so as to synchronize operations. For example, when a batch of parts is being produced, an urgent order crops up, the production manager goes to adjust the schedule, the warehouse prepare the material, and the logistics initiate a shipping task accordingly via each application. Hence, the next production job could be launched as punctually as possible.

Synchroperation in manufacturing anticipates high demand for a traceable and visible system that can synchronize the cyber and physical world greatly regarding organization, data, and decision. For the cyber-physical synchronization, IoT technologies play an essential role in digitizing physical objects to provide a foundation for information traceability and visibility. For the spatio-temporal synchronization, objects in operations are highly intertwined concerning time and space since the flow of men, machines, and materials in the factory are much frequent and intricate. To model and visualize these operations, a spatio-temporal analytics method is designed to segment space and time of operations for local dissolution and then integrate them to reap the overall effects. For the data-driven decision synchronization, consortium blockchain provides an effective solution to sharing information among the platform safely and helping users easily track and trace. Based on data analytics engines, different applications are developed for data visualization to assist operators in making decisions. In this case, the visibility and traceability of the system get considerably enhanced.

5. Graduation Intelligent Manufacturing System for Synchroperation

This section proposes the GiMS with “divide and conquer” principles, to address the complex, stochastic, and dynamic nature of manufacturing for achieving synchroperation. The basic form and principles of GiMS can be found in previous research (Lin, et al., 2019; Guo et al., 2020a). As shown in Fig. 3, this paper presents the five key phases to implement GiMS in factories.

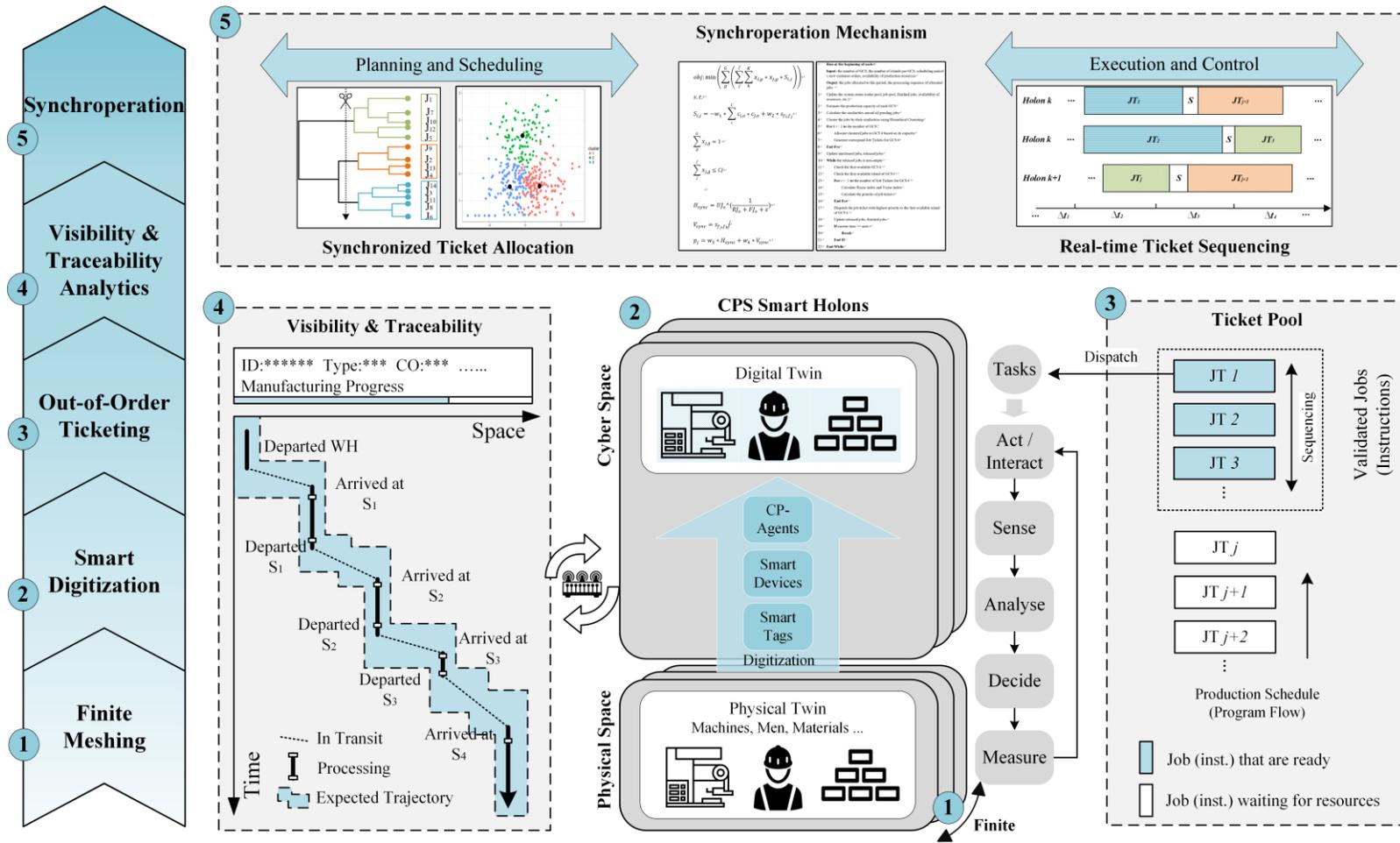


Fig. 3. Five phases of GiMS

Phase 1: Finite Meshing

This phase involves spatially dividing the factory organization and temporally discretizing the decision horizon into an equivalent system of finite “Graduation Ceremony Stages (GCS, a space unit in a time period)”, to minimize complexity and localize uncertainties. Operation elements (men, machines, materials) are defined for all “stages” as physical twins.

The space scope of a factory contains various production, logistics, and storage facilities and areas that can be divided into smaller space units. The decision horizon is discretized into multiple shorter time periods (Balakrishnan and Cheng, 2007; Torkaman et al., 2017). Because the space unit and time period of GCS in meshing are small enough relative to the original system, the subproblem size is limited, and a straightforward decision model can be built. All the uncertainties that occur during the current period can be considered in the next period with negligible loss of service quality. There are different rules to generate the mesh: (1) Meshing according to absolute space and time. Usually, the factory is divided into finite space units based on their absolute spatial positions; the decision horizon can be discretized into multiple time periods representing a working shift or several hours. (2) Meshing based on functionality. This rule divides the factory into GCS with different functions. For instance, logistics facilities and production workstations belong to separate GCS. Production area can be further divided based on their functions and responsibilities (threading, heat-treating, etc.) to obtain finer granularity. Correspondingly, the time scope should be discretized with the characteristics of different function areas taken into account. (3) Meshing on a dynamic basis. This rule dynamically adjusts the granularity of spatiotemporal mesh based on the real-time situation. This usually requires a high level of visibility and traceability throughout the fast-changing supply chain.

Phase 2: Smart Digitization

Phase 2 is to digitize the operation elements at all GCS for generating digital twins with the enabling technologies. A highly visible, transparent, and interconnected Cyber-Physical Factory (CPF) with real-time visibility and traceability is built through smart digitization. This phase constructs the data dimension of physical twins, and the cyber-physical synchronization is achieved through smart gateways.

With the deployment of mobile crowdsensing, IIoT devices, and cyber-physical agents, all physical entities in the factory are digitized for generating cyber avatars. The physical twins combine with corresponding digital twins to form CPS smart holons (CPS-SHs) that are decentralized and of autonomy to some extent. All CPS-SHs are physically independent but digitally interconnected by the tasks. The capabilities of CPS-SHs include: 1) sense the environment, such as how it connects with other holons and the status of task pool; 2) analyze the real-time production data and information; 3) autonomously make decisions based on the status of tasks and the state of itself; 4) take actions accordingly and interact with each other; 5) finally measure the key production performance (e.g., holon utilization, efficiency). Relevant applications, services, and analytics are integrated into the cyber space. Smart gateways are deployed to synchronize cyber and physical spaces, analyze the status of smart holons, and support visibility and traceability analytic between smart holons.

Phase 3: Out-of-Order Ticketing

This phase implements an Out-of-Order (OoO) ticketing for smart holons to facilitate smooth onsite execution and flexible control of production progress with enhanced resilience. OoO ticketing guarantees the data-driven decision synchronization at the operational level.

Three kinds of tickets, including job ticket (JT), setup ticket (ST), operation ticket (OT) and twined logistics ticket (LT) are critical in GiMS. Smart tags serve as the carriers of digitalized tickets. JTs are designed for permitting the right jobs to produce in one batch considering demand and capacity. STs are designed to control flexible setups between different job families so that the setup can be informed in advance and performed at the right time. OTs and twined LTs are designed to synchronize operation and JIT delivery. These tickets are real-timely generated and allocated to the ticket pools of each GCS. OoO is a paradigm used in modern CPUs to avoid stalls and improve processing efficiency (Hwu & Patt, 1986). OoO allows the processor to execute instructions in an order governed by the availability of input data and execution units. By analogy, the OoO ticketing in factories allows jobs to be processed in an order governed by the availability of materials, machines, and men. That is, smart holons look ahead in the ticket pool and find those that are ready to be processed. When a disturbance like material deficiency/loss or machine failure occurs, the

smart holon can decide to process other ready jobs or using other available machines rather than wait. OoO ticketing organizes the onsite production operations with simplicity and resilience, and offer a robust and straightforward logic to tackle frequent uncertainties in the real-life shop floor.

Phase 4: Visibility and Traceability Analytics

In phase 4, the cyber-physical visibility and traceability (CPVT) analytics is utilized to identify and establish the dependencies and connectivity of GCS and smart holons, and to mitigate the spatiotemporal uncertainties. Besides, this phase also serves as the foundation for spatio-temporal synchronization and higher-level data-driven decision synchronization.

The holonic dependency and connectivity usually refer to the logical relationship between holons, such as how the state of ticket pools update over time, and how the tickets flow between holons. The CPVT is the key tool to real-timely monitor ticket pools and to establish the connectivity from two aspects. Firstly, how the states of holons update over time: The input of one holon at the beginning of the current time unit consists of two parts. The first part is the output of that holon in the previous time unit; the second part contains new information that occurs in the previous time unit; these two parts are also strongly influenced by various uncertainties such as stochastic processing time, machine failure, absence of operator, etc. Secondly, how the tickets flow between holons: Completing a production job usually requires performing several operations. These operations and their intrinsic sequencing and spatial constraints are defined in the tickets. The jobs whose operation at the current holon in the previous time unit has been completed, will be transferred to the following holons. And the jobs transferred to the current holon in the current time unit, are the union set of jobs whose operation in the previous holon has been completed. These real-time data and information are vital for supporting decision-making within GCS and connecting all holons.

Phase 5: Synchroperation

This phase designs the synchroperation mechanism under GiMS to facilitate upper-level planning and scheduling and lower-level onsite execution and control. Spatio-temporal synchronization and data-driven decision synchronization are achieved in this phase.

In the upper-level planning and scheduling, the overall planning horizon T is discretized into multiple shorter scheduling periods t . Based on the real-time demand (e.g., product type, quantity) and production constraints (e.g., capacity, resource) in period t , the job ticket allocation mechanism aims to generate schedule for period $t + 1$ on an aggregate basis for families of jobs and allocate job tickets. The similarity among jobs is usually measured from the aspects of setup, material requirement, operator skill requirement, correlative orders and so on. In the lower-level execution and control, as the jobs allocated to a single period are similar, the rigid sequence of these jobs is less significant. Thus, OoO ticketing of tickets is adopted. At the beginning of t , job tickets are released to each GCS task pool, the operation tickets and logistics tickets for this job are activated. A job ticket is validated once all required resources are available. When there is a vacancy in the workstation buffer, and the priorities of validated tickets are calculated based on real-time data, the job ticket with the highest priority will be dispatched. The priorities are usually computed based on horizontal synchronization, vertical synchronization, and the matching degree of the job and workstation (job type, machine type, operator skill etc.). The bi-level synchroperation mechanism promises both optimized decisions at the managerial level and resilience execution, flexible control at the operational level.

6. Case study: Synchroperable hybrid assembly line Based on GiMS

In this section, the GiMS is applied to a novel manufacturing layout named hybrid assembly line (HAL). This layout is adopted by a world-leading air conditioner manufacturer G to face the fast-changing market with high flexibility. First, the GiMS enabled HAL is introduced, and then a comprehensive numerical analysis is carried out to verify the effectiveness of the proposed method. Results show that GiMS can obtain significant performance improvements regarding synchroperability measures.

6.1. GiMS enabled HAL

Fig. 4 presents the GiMS-enabled HAL. The HAL consists of a series of assembly stations. Each station has the space for equipment, materials and tools. All operations for one product will be processed and completed on an assembly trolley. The operators move the assembly trolley

along the assembly line for assembly operations. After handling the final product, the operator moves the assembly trolley back to the first station for next assembly job. In the assembly process, the job finished early can overtake the ones in front. For example, in the real time job pool part of the Fig. 4, the job #3 is completed faster in station 1, and it will become the first job in next station. Operators in the same station share the same job pool and always take the first job from the job pool. Each operator can only process one job at a time and preemption of jobs is not allowed.

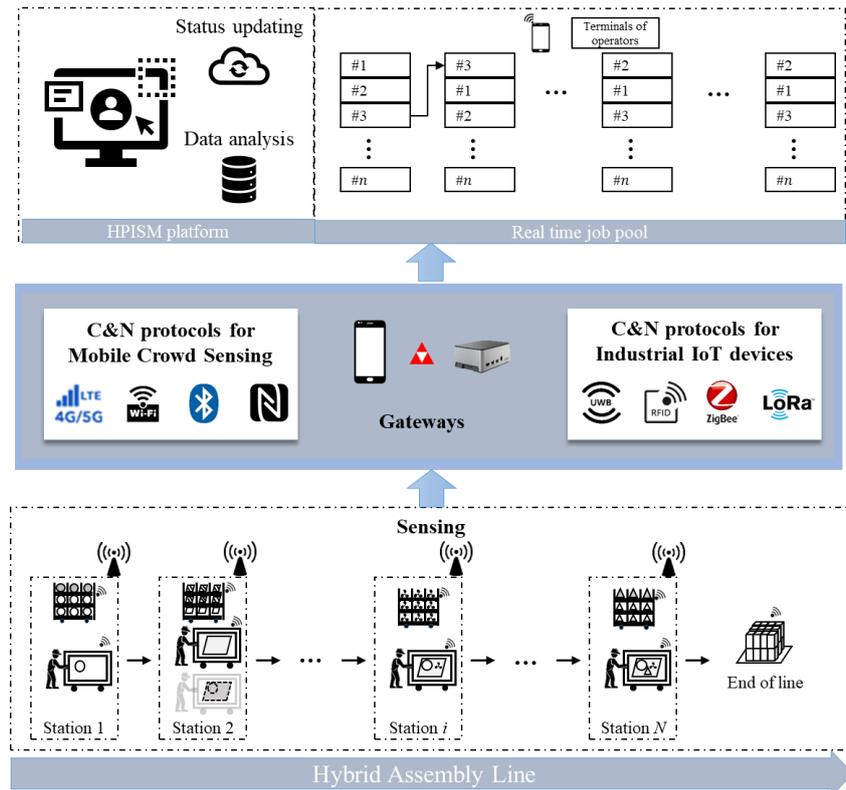


Fig. 4. GiMS enabled HAL assembly process

The PI infrastructure is deployed for creating the hyperconnected cyber-physical manufacturing environment. The production data and information are real-timely captured and transmitted to cyberspace through smart gateways with cyber-physical visibility and traceability. The whole factory is meshed spatially and temporally with the proposed meshing rules in section 5. In this process, each HAL is regarded as a GCS and there are multiple homogeneous GCSs in the factory. From the temporal perspective, different strategies including integral time units such as 1 hour or the average processing time can be applied. Smart tags (e.g., RFID tags) are attached to the corresponding machine, materials, and tools, and all elements are digitalized for generating cyber avatars. The CPS-SHs are formed from cyber avatars and their digital twins. Finally, the

entire factory is discretized into CPS-SHs.

The job allocation and execution process under HAL is formulated mathematically to ensure the synchroperation in the whole assembly process. The customer order o_i is denoted by

$$o_i \triangleq (at_i, d_i, n_{i,p})$$

The at_i and d_i represent the arrival time and due date of the i th order, and $n_{i,p}$ donates the required amount of product type p . Each product is represented by an assembly job ticket. The assembly job tickets are released dynamically based on the comprehensive priority (CP). The CP_j of job j is calculated with V-sync and H-sync proposed in GIMS and due date priority (DP):

$$CP_j = w_H \cdot norm(HS_j) + w_V \cdot norm(VS_j) + w_D \cdot norm(DP_j) \quad (1)$$

$$HS_j = \left(\frac{UJ_i}{\alpha} \right)^{\frac{1}{(RJ_i/\alpha)^{\beta}}} \quad (2)$$

$$VS_j = \frac{b_{p_j, p_l}}{1 + \prod_{l' \in L-l} b_{p_j, p_{l'}}} \quad (3)$$

$$DP_j = d_i - t \quad (4)$$

The job with smaller CP_j has higher priority. HS_j represents the production progress of the order that contains job j . UJ_i and RJ_i denote the unreleased jobs and released jobs of order i respectively, ε is an arbitrary small real number in case RJ_i is equal to 0. α and β are positive real number used to smooth the HS_j value. As setup is required for changeover between different product types, VS_j reflects the matching degree of the job j and the available line l (whether setup is required). p_j and p_l denote the product type of job j and the setup condition of line l left by previous job. b_{p_j, p_l} is a Boolean function and takes the following form:

$$b_{p_j, p_l} = \begin{cases} 0 & \text{if } p_j = p_l \\ 1 & \text{otherwise} \end{cases} \quad (5)$$

For DP_j , d_i and t are the modified due date of order i and current time. The normalization function takes the following form:

$$norm(x) = \frac{x - \min}{\max - \min} \quad (6)$$

Where the maximum and minimum can be derived from all unrealised jobs in the system. Then, the objective in the release process is to allocate the jobs with higher priorities

$$\text{Minimize } \sum_{j=1}^J CP_j x_j$$

s.t.

$$\sum_{j=1}^J x_j \leq N \quad (7)$$

$$x_j \in \{0,1\} \quad (8)$$

N represents the maximum number of jobs released to each HAL and depends on the operators in each station, time period t and average job processing time. Then, a batch of jobs with higher priorities will be released to the job pool of the first station in the HAL. Then, in each assembly line, assembly job tickets are handled under the OoO ticketing. Notations used in the system are summarized in Table 3.

Table 3. Notations used in the system

Index	
i	customer index
j	job index
l	HAL index
Parameters	
I	total customer orders
J	all unreleased assembly jobs
t	time period
o_i	customer order i .
d_i	order i 's due date.
$n_{i,p}$	the required amount of product type p in order i .
CP_j	comprehensive priority of job j .
HS_j	H-sync priority of job j .
UJ_i	unreleased jobs of order i .
RJ_i	released jobs of order i .
VS_j	V-sync priority of job j .
DP_j	due date priority of job j .
α, β	positive real numbers.
p_j	denote the product type of job j and the setup condition of line l
p_l	the setup condition of line l .
N	the maximum number of released jobs in each releasing decision.
c_j	the complete time of job j
r_j	released time of job j
TST_l	the total setup time of line l .
Decision variable	
x_j	if job j is released

6.2. Numerical Study

Several experiments are conducted to evaluate the performance of GIMS in the HAL case. Several synchronizability measures are used in the experiments. For the simultaneity measure, the average flow time (AFT) is adopted, which considers the time duration from the first assembly job starts processing on the HAL to completion of all jobs. The AFT is calculated by

$$AFT = \frac{\sum_{i=1}^I \left(\max_{j \in o_i} c_j - \min_{j \in o_i} r_j \right)}{I} \quad (9)$$

Where c_j and r_j represent the complete time and released time of job j . The average tardiness (ATD) is used as the punctuality measure and takes the following form:

$$ATD = \frac{\sum_{i=1}^I \left(\max \left(0, \max_{j \in o_i} c_j - d_i \right) \right)}{I} \quad (10)$$

The makespan (MS) and average setup time (AST) are adopted as the cost efficiency measure of the whole plant. The AST can be calculated as follows:

$$AST = \frac{\sum_{l=1}^L (TST_l)}{L} \quad (11)$$

TST_l represents the total setup time of line l . The parameters for generating test instances are set as the values in Table 4.

Table 4. Experimental data

Parameters	Value
Cell	5
HALs	5
Operators	4
Average processing time in each station (min)	$N(10,2)$
Average setup time (min)	10,20, 30, 40
Customer orders	50
Number of jobs per order	5
Average orders inter-arrival time (min)	15,25,35
Due dates of orders	$U [7, 13] \times \text{order inter-arrival time}$
Products	3, 6, 9,12
Released job numbers	4
Time period	100

The operation time for each job in each station follows a normal distribution with mean 10 and standard deviation 2. The number of customer orders is set to 50. The time period, setup time, and types of products are set to several fixed numbers. The order inter-arrival time is generated from a Poisson process with mean 15, 25 and 35. The inter-arrival time which balances the output

rate and job arrival rate of the entire plant. When the value becomes smaller, which means that orders come to the system more frequently and can be considered as the peak season. Similarly, a smaller value means the offseason. The type of products are set to {3, 6, 9, 12} respectively. The due date of each order is set to 7~13 times of average orders inter-arrival time (min) after the order arrives. 4 job are released each time and the time period is set to 100.

The performance of GiMS is compared to several common scheduling rules with different numbers of orders. Then the sensitivity analysis is conducted to investigate the effects of two crucial factors: product type and setup time. To compare the performance, the following three common rules are adopted: first-come-first-serve (FCFS) rule (Schwiegelshohn and Yahyapour, 1998), Shortest processing time (SPT) (Pickardt and Branke, 2012) and Earliest due date (EDD) (Baker, 1984). In this case, SPT considers that the order with less unreleased jobs has higher priority. The setup time is set to 30 and product type is set to 6.

At the beginning of the production horizon, each HAL holds the same number of jobs with the same product type. The inter-arrival time is gradually increased from 15 to 35 in steps of 10 and there are 50 customer orders. Under each number of orders, the experiments for the four rules are carried out individually for 100 times, which is large enough to give statistically reliable results. Table 5 shows the average values of the results.

Table 5. Synchroperability measures performance of HALs under different rules

Inter-arrival time	Rules	Measures			
		AFT	ATD	MS	TST
15	GIMS	217.3851	72.3005	1216.085	1785.572
	FCFS	221.9514	76.75619	1371.398	2157.09
	SPT	221.3701	76.37134	1363.869	2156.957
	EDD	221.8738	76.64523	1372.187	2157.06
25	GIMS	211.2348	5.708845	1208.925	1824.576
	FCFS	219.4813	7.7747	1596.651	2204.133
	SPT	219.8993	7.905285	1575.59	2203.765
	EDD	219.6664	7.769687	1567.685	2203.273
35	GIMS	208.4382	0.005619	1506.978	1827.535
	FCFS	214.5554	0.173307	1586.747	2171.051
	SPT	214.3138	0.13862	1583.864	2171.605
	EDD	214.6878	0.146956	1602.94	2171.842

Numerical results show that GIMS achieves lower values in all measures. Especially, GIMS

achieves a 15%~18% reduction in TST compared to other rules. This means that GiMS can effectively reduce the setup operations can help the manufacturer achieve cost-efficiency and higher synchronoperability. Another key observation is that GiMS achieve less ATD when the inter-arrival time is 15 and enable orders to be finished simultaneously. This means that during peak seasons, GiMS can reduce setup cost without sacrificing the punctuality performance.

The sensitivity analysis is conducted to the product type and setup time. First, the product type is increased from 3 to 12 in steps of 3 with the setup time is 30. Also, FCFS policy is used for comparison. The results are shown in Fig. 5. And then the setup value is increased from 10 to 40 in steps of 10 with 6 product types. The results are shown in Fig. 6.

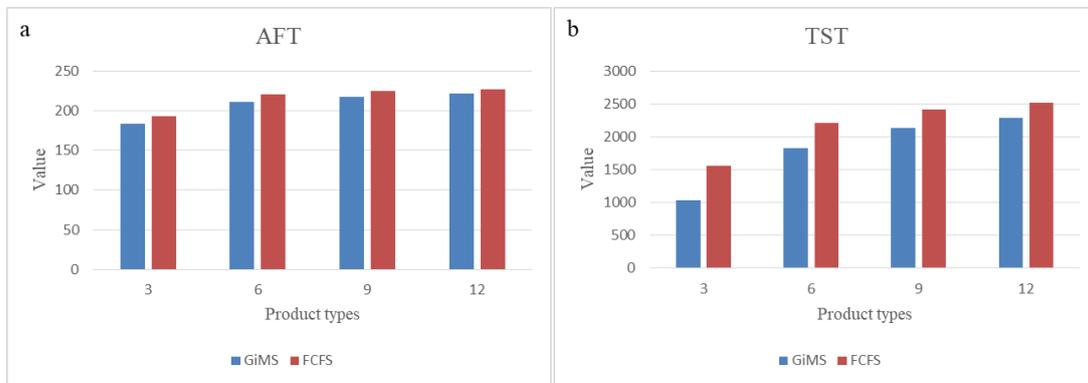


Fig. 5. Measures of GiMS and FCFS policy under different product types

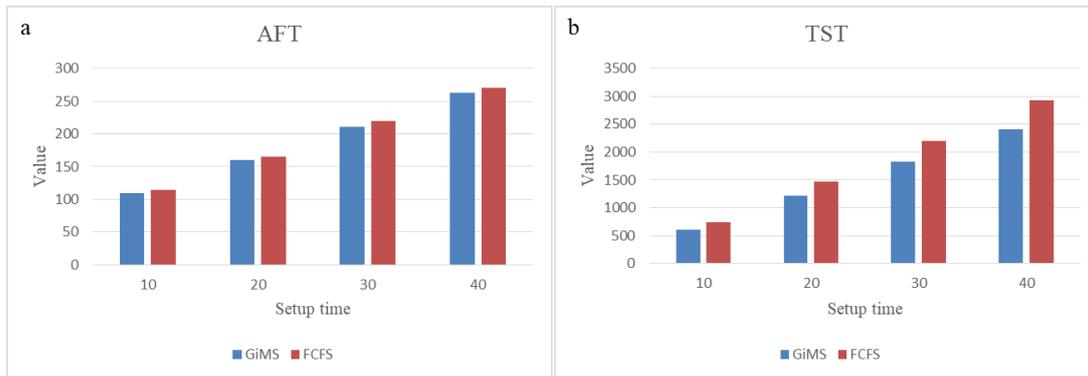


Fig. 6. Measures of GiMS and FCFS policy under different setup times

Fig. 5 shows that GiMS has less AFT and TST compared to FCFS policy in each instance. This implies that GiMS can achieve better performance in a wide range of regimes. When product types is 3 and 6, GiMS outperforms FCFS in AFT. For TST, it can be seen that GiMS also has a better performance in the case with less product types. In Fig. 6, GiMS has less AFT and TST compared to FCFS policy under each setup time. For AFT and TST, GiMS has a greater advantage

over FCFS when the setup time increases. Besides, the TST has been significantly reduced compared to AFT. This indicates that GiMS can help manufacturers to reduce the setup time significantly especially when the setup time is larger.

The management insights can be summarized as follows: GiMS can achieve higher synchronoperability and significant performance improvement for HAL. With the deployment of HPISMP, real-time status of operators, equipment, and materials is available for decision making. This enables a global optimization of decisions.

7. Conclusions and Future perspectives

Industry 4.0 connotes a new industrial revolution with the convergence between physical and digital spaces, which are currently revolutionizing the way that production operations are managed. To explore the evolution of production and operations management paradigms in the era of Industry 4.0, a concept of manufacturing synchronoperation with enabling technologies and associated methodologies are proposed for transformation and implementation of Industry 4.0 manufacturing.

The main contributions of this paper can be concluded as follows: (1) A concept of manufacturing synchronoperation with cyber-physical synchronization, data-driven decision synchronization and spatio-temporal synchronization, is proposed for Industry 4.0 production and operations management. (2) A HPISMP assisted with digital twin and consortium blockchain is developed as a technical solution to support the transformation of manufacturing synchronoperation. (3) GiMS with “divide and conquer” principles is proposed as a methodology to address the complex, stochastic, and dynamic nature of manufacturing for achieving synchronoperation. (4) The potential advantages of implementation of manufacturing synchronoperation are illustrated with an industrial case from an air conditioner manufacturer.

This paper presents a new paradigm of production and operations management in the era of Industry 4.0-manufacturing synchronoperation. The research is still in its infancy, and there are abundant research opportunities in this topic. Further research efforts on principles, methodologies, and support technologies for transforming production and operations management to Industry 4.0 manufacturing are necessary. Several possible research directions with related research questions

are listed as follows.

RQ1: How manufacturing synchroperation reshapes the way manufacturer do business with their customers? How to establish adaptive business models for manufacturing synchroperation in the era of Industry 4.0?

Industry 4.0 manufacturing is revolutionizing the way that production operations are managed and done, which also has the potential to revolutionize the way manufacturer do business with their customers and suppliers. The concept of manufacturing synchroperation provides an insight for manufacturers to re-evaluate and develop their business model to capture and maximize the value of the customer in the Industry 4.0 manufacturing environment. More innovative business models need to be further explored.

RQ2: How to measure the disruptions of manufacturing synchroperation on supply chain? How to integrate manufacturing synchroperation with the processes of supply, warehousing and delivery to increase the agility of the supply chain?

Manufacturing synchroperation with cyber-physical synchronization promises to remove information or communications barriers cross multi-echelon and inter-organizational activities. Effective methods to measure the disruptions of manufacturing synchroperation on supply, warehousing and delivery processes, and increase the agility of the whole supply chain through the real-time cyber-physical visibility and traceability deserve further explorations.

RQ3: What is the technical requirement for transforming to Industry 4.0 manufacturing? How to design effective technical standards and architectures for real-time information exchange among "real-time things" to support manufacturing synchroperation?

The transformation of Industry 4.0 manufacturing requires technical infrastructures to support real-time visibility and information sharing. Technical standards and architectures with a high degree of connectivity, interoperability and accessibility must be designed to define the specifications for real-time information exchange among "real-time things", which is the basis for achieving manufacturing synchroperation in the era of Industry 4.0.

RQ4: What are the effects of the real-time visibility and information sharing on complexity and uncertainty nature of manufacturing? How to model and minimize the

uncertainty and complexity in real-time manufacturing environment?

Although the potential benefits of the real-time visibility and information sharing in the era of Industry 4.0 have been acknowledged in general, the theoretical foundations are rarely considered. Innovative methods to model and measure the effects of the real-time visibility and information sharing on complexity and uncertainty nature of manufacturing, and minimize the uncertainty and complexity in the Industry 4.0 manufacturing environment are crucial.

RQ5: How to realize the full potentials of historical and real-time production data? How to derive decentralized/autonomous decision-making to create data-driven value-adding services for Industry 4.0 manufacturing?

The hyper-connection, digitization and sharing in Industry 4.0 manufacturing bring new decentralized production patterns with real-time production data. Under this circumstance, decentralized elements in the production system can independently or collaboratively make decisions and even take actions. Therefore, how to derive decentralized/autonomous decision-making from the fusion of enormous production data and corresponding management strategies to create data-driven value-adding services need to be investigated.

Acknowledgement

Acknowledgement to Zhejiang Provincial, Hangzhou Municipal, Lin'an City Governments, Hong Kong ITF Innovation and Technology Support Program (ITP/079/16LP) and financial support from the 2019 Guangdong Special Support Talent Program-Innovation and Entrepreneurship Leading Team (China) (2019BT02S593).

Reference

- Akkermans, H.A., van der Horst, H., 2002. Managing IT infrastructure standardisation in the networked manufacturing firm. *International Journal of Production Economics*, 75, 213-228.
- Balakrishnan, J., & Cheng, C. H., 2007. Multi-period planning and uncertainty issues in stationular manufacturing: A review and future directions. *European Journal of Operational Research*, 177(1), 281-309.
- Baker, K. R., 1984. Sequencing rules and due-date assignments in a job shop. *Management Science*, 30(9), 1093-1104.

- Browne, J., Dubois, D., Rathmill, K., Sethi, S.P., Stecke, K.E., 1984. Classification of flexible manufacturing systems. *The FMS magazine*, 2, 114-117.
- Buzacott, J. A., & Yao, D. D., 1986. Flexible manufacturing systems: a review of analytical models. *Management Science*, 32(7), 890-905.
- Chen, J., Huang, G. Q., Luo, H., & Wang, J., 2015. Synchronisation of production scheduling and shipment in an assembly flowshop. *International Journal of Production Research*, 53(9), 2787-2802.
- Chen, J., Wang, M., Kong, X. T., Huang, G. Q., Dai, Q., & Shi, G., 2019. Manufacturing synchronization in a hybrid flowshop with dynamic order arrivals. *Journal of Intelligent Manufacturing*, 30(7), 2659-2668.
- Chen, T., Chiu, M.C., 2017. Development of a cloud-based factory simulation system for enabling ubiquitous factory simulation. *Robotics and Computer-Integrated Manufacturing*, 45, 133-143.
- D'Amours, S., Montreuil, B., Lefrancois, P., Soumis, F., 1999. Networked manufacturing: The impact of information sharing. *International Journal of Production Economics*, 58, 63-79.
- Da Xu, L., He, W., & Li, S., 2014. Internet of things in industries: A survey. *IEEE Transactions on Industrial Informatics*, 10(4), 2233-2243.
- Fang, J., Qu, T., Li, Z., Xu, G. and Huang, G.Q., 2013. Agent-based gateway operating system for RFID-enabled ubiquitous manufacturing enterprise. *Robotics and Computer-Integrated Manufacturing*, 29(4), pp.222-231.
- Fazlollahtabar, H., Saidi-Mehrabad, M., & Balakrishnan, J., 2015. Mathematical optimization for earliness/tardiness minimization in a multiple automated guided vehicle manufacturing system via integrated heuristic algorithms. *Robotics and Autonomous Systems*, 72, 131-138.
- Ganti, R.K., Ye, F. and Lei, H., 2011. Mobile crowdsensing: current state and future challenges. *IEEE communications Magazine*, 49(11), pp.32-39.
- GE, 2020. Predix Platform: Connect, optimize, and scale your digital industrial applications. Available online: <https://www.ge.com/digital/iiot-platform> (accessed on 22 October 2020).
- Gunasekaran, A., 1999. Agile manufacturing: A framework for research and development. *International Journal of Production Economics*, 62, 87-105.
- Guo, D., Li, M., Zhong, R. and Huang, G.Q., 2020a. Graduation Intelligent Manufacturing System (GiMS): an Industry 4.0 paradigm for production and operations management. *Industrial Management & Data Systems*.
- Guo, D., Zhong, R. Y., Rong, Y., and Huang, G. Q., 2020d. Synchronization between manufacturing and logistics under IIoT and digital twin-enabled Graduation Intelligent Manufacturing System. *IEEE Transactions on Industrial Informatics*, under review.
- Guo, D., Zhong, R. Y., Lin, P., Lyu, Z., Rong, Y., & Huang, G. Q., 2020c. Digital twin-enabled Graduation Intelligent Manufacturing System for fixed-position assembly islands. *Robotics and Computer-Integrated Manufacturing*, 63, 101917.
- Guo, D., Zhong, R. Y., Ling, S., Rong, Y., & Huang, G. Q., 2020d. A roadmap for Assembly 4.0: self-configuration of fixed-position assembly islands under Graduation Intelligent Manufacturing System. *International Journal of Production Research*, 58(15), 4631-4646.
- Hsu, S. Y., & Liu, C. H., 2009. Improving the delivery efficiency of the customer order scheduling

- problem in a job shop. *Computers & Industrial Engineering*, 57(3), 856-866.
- Huang, G.Q., Wright, P.K., Newman, S.T., 2009. Wireless manufacturing: a literature review, recent developments, and case studies. *International Journal of Computer Integrated Manufacturing*, 22, 579-594.
- Huang, G.Q., Zhang, Y.F., Jiang, P.Y., 2008. RFID-based wireless manufacturing for real-time management of job shop WIP inventories. *The International Journal of Advanced Manufacturing Technology*, 36, 752-764.
- Hwu, W.-r., & Patt, Y. N., 1986. Hpsm, a high performance restricted data flow architecture having minimal functionality. *ACM SIGARCH Computer Architecture News*, 14(2), 297-306.
- Ivanov, D., Tang, C. S., Dolgui, A., Battini, D., & Das, A., 2020. Researchers' perspectives on Industry 4.0: multi-disciplinary analysis and opportunities for operations management. *International Journal of Production Research*, 1-24.
- Koh, L., Orzes, G., & Jia, F. J., 2019. The fourth industrial revolution (Industry 4.0): technologies disruption on operations and supply chain management. *International Journal of Operations & Production Management*, 39, 817-828.
- Kong, X.T., Zhong, R.Y., Zhao, Z., Shao, S., Li, M., Lin, P., Chen, Y., Wu, W., Shen, L., Yu, Y. and Huang, G.Q., 2020. Cyber physical ecommerce logistics system: An implementation case in Hong Kong. *Computers & Industrial Engineering*, 139, p.106170.
- Kortuem, G., Kawsar, F., Sundramoorthy, V., & Fitton, D., 2009. Smart objects as building blocks for the internet of things. *IEEE Internet Computing*, 14(1), 44-51.
- Kusiak, A., 2017. Smart manufacturing must embrace big data. *Nature*, 544(7648), 23-25.
- Lee, J., Bagheri, B., & Kao, H. A., 2015. A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manufacturing Letters*, 3, 18-23.
- Li, Z., Barenji, A.V. and Huang, G.Q., 2018. Toward a blockchain cloud manufacturing system as a peer to peer distributed network platform. *Robotics and Computer-Integrated Manufacturing*, 54, pp.133-144.
- Lin, P., Li, M., Kong, X., Chen, J., Huang, G. Q., & Wang, M., 2018. Synchronisation for smart factory-towards IoT-enabled mechanisms. *International Journal of Computer Integrated Manufacturing*, 31(7), 624-635.
- Lin, P., Shen, L., Zhao, Z., & Huang, G. Q., 2019. Graduation manufacturing system: synchronization with IoT-enabled smart tickets. *Journal of Intelligent Manufacturing*, 30(8), 2885-2900.
- Lin, Y. C., & Chen, T., 2017. A ubiquitous manufacturing network system. *Robotics and Computer-Integrated Manufacturing*, 45, 157-167.
- Luo, H., Wang, K., Kong, X. T., Lu, S., & Qu, T., 2017. Synchronized production and logistics via ubiquitous computing technology. *Robotics and Computer-Integrated Manufacturing*, 45, 99-115.
- Luo, H., Yang, X., & Wang, K., 2019. Synchronized scheduling of make to order plant and cross-docking warehouse. *Computers & Industrial Engineering*, 138, 106108.
- Martinez, M.T., Fouletier, P., Park, K.H., Favrel, J., 2001. Virtual enterprise-organisation, evolution and control. *International Journal of Production Economics*, 74, 225-238.
- McFarlane, D., Sarma, S., Chirn, J.L., Wong, C.Y., Ashton, K., 2003. Auto ID systems and

- intelligent manufacturing control. *Engineering Applications of Artificial Intelligence*, 16, 365-376.
- McGehee, J., Hebley, J., & Mahaffey, J., 1994. The MMST computer-integrated manufacturing system framework. *IEEE Transactions on Semiconductor Manufacturing*, 7(2), 107-116.
- Montreuil, B., Frayret, J.M., D'Amours, S., 2000. A strategic framework for networked manufacturing. *Computers in industry*, 42, 299-317.
- Mourad, M.H., Nassehi, A., Schaefer, D., Newman, S.T., 2020. Assessment of interoperability in cloud manufacturing. *Robotics and Computer-Integrated Manufacturing*, 61.
- Nagalingam, S.V., Lin, G.C.I., 1999. Latest developments in CIM. *Robotics and Computer-Integrated Manufacturing*, 15, 423-430.
- Newman, S.T., Nassehi, A., Xu, X.W., Rosso, R.S.U., Wang, L., Yusof, Y., Ali, L., Liu, R., Zheng, L.Y., Kumar, S., Vichare, R., Dhokia, V., 2008. Strategic advantages of interoperability for global manufacturing using CNC technology. *Robotics and Computer-Integrated Manufacturing*, 24, 699-708.
- Olsen, T. L., & Tomlin, B., 2020. Industry 4.0: opportunities and challenges for operations management. *Manufacturing & Service Operations Management*, 22(1), 113-122.
- Pickardt, C. W., & Branke, J., 2012. Setup-oriented dispatching rules—a survey. *International Journal of Production Research*, 50(20), 5823-5842.
- Qu, T., Lei, S.P., Wang, Z.Z., Nie, D.X., Chen, X., Huang, G.Q., 2016. IoT-based real-time production logistics synchronization system under smart cloud manufacturing. *The International Journal of Advanced Manufacturing Technology*, 84, 147-164.
- SAP, 2020. SAP Cloud Platform. Available online: <https://www.sap.com/hk/products/cloud-platform.html> (accessed on 22 October 2020).
- Schwiegelshohn, U., & Yahyapour, 1998. Analysis of first-come-first-serve parallel job scheduling. *SODA*, 98, 629-638.
- Siemens, 2020. MindSphere: Connecting the things that run the world. Available online: <https://siemens.mindsphere.io/en> (accessed on 22 October 2020).
- Stecke, K.E., 1983. Formulation and Solution of Nonlinear Integer Production Planning Problems for Flexible Manufacturing Systems. *Management Science*, 29, 273-288.
- Tao, F., Cheng, J., Qi, Q., Zhang, M., Zhang, H., & Sui, F., 2018. Digital twin-driven product design, manufacturing and service with big data. *The International Journal of Advanced Manufacturing Technology*, 94(9-12), 3563-3576.
- Torkaman, S., Ghomi, S. F., & Karimi, B., 2017. Multi-stage multi-product multi-period production planning with sequence-dependent setups in closed-loop supply chain. *Computers & Industrial Engineering*, 113, 602-613.
- Wang, X., Ong, S.K., Nee, A.Y.C., 2018a. A comprehensive survey of ubiquitous manufacturing research. *International Journal of Production Research*, 56, 604-628.
- Wang, X.V., Wang, L.H., Gordes, R., 2018b. Interoperability in cloud manufacturing: a case study on private cloud structure for SMEs. *International Journal of Computer Integrated Manufacturing*, 31, 653-663.
- Wang, X.V., Wang, L.H., Mohammed, A., Givehchi, M., 2017. Ubiquitous manufacturing system based on Cloud: A robotics application. *Robotics and Computer-Integrated Manufacturing*,

45, 116-125.

- Wang, X.V., Xu, X.W., 2013. An interoperable solution for Cloud manufacturing. *Robotics and Computer-Integrated Manufacturing*, 29, 232-247.
- Xu, G., Wang, J., Huang, G.Q. and Chen, C.H., 2017. Data-driven resilient fleet management for cloud asset-enabled urban flood control. *IEEE Transactions on Intelligent Transportation Systems*, 19(6), pp.1827-1838.
- Xu, X., 2012. From cloud computing to cloud manufacturing. *Robotics and Computer-integrated Manufacturing*, 28(1), 75-86.
- Xu, X.W., Wang, H., Mao, J., Newman, S.T., Kramer, T.R., Proctor, F.M., Michaloski, J.L., 2005. STEP-compliant NC research: the search for intelligent CAD/CAPP/CAM CNC integration. *International Journal of Production Research*, 43, 3703-3743.
- Yin, Y., Stecke, K. E., & Li, D., 2018. The evolution of production systems from Industry 2.0 through Industry 4.0. *International Journal of Production Research*, 56(1-2), 848-861.
- Yusuf, Y.Y., Sarhadi, M., Gunasekaran, A., 1999. Agile manufacturing: The drivers, concepts and attributes. *International Journal of Production Economics*, 62, 33-43.
- Zhang, L., Luo, Y.L., Tao, F., Li, B.H., Ren, L., Zhang, X.S., Guo, H., Cheng, Y., Hu, A.R., Liu, Y.K., 2014. Cloud manufacturing: a new manufacturing paradigm. *Enterprise Information Systems*, 8, 167-187.
- Zhang, Y., Qu, T., Ho, O.K. and Huang, G.Q., 2011. Agent-based smart gateway for RFID-enabled real-time wireless manufacturing. *International Journal of Production Research*, 49(5), pp.1337-1352.
- Zhang, Y.F., Qu, T., Ho, O., Huang, G.Q., 2011. Real-time work-in-progress management for smart object-enabled ubiquitous shop-floor environment. *International Journal of Computer Integrated Manufacturing*, 24, 431-445.
- Zhao, Z., Fang, J., Huang, G.Q. and Zhang, M., 2017. Location management of cloud forklifts in finished product warehouse. *International Journal of Intelligent Systems*, 32(4), pp.342-370.
- Zhong, R.Y., Dai, Q.Y., Qu, T., Hu, G.J., Huang, G.Q., 2013. RFID-enabled real-time manufacturing execution system for mass-customization production. *Robotics and Computer-Integrated Manufacturing*, 29, 283-292.
- Zhong, R.Y., Xu, X., Klotz, E., Newman, S.T., 2017. Intelligent Manufacturing in the Context of Industry 4.0: A Review. *Engineering*, 3, 616-630.

Graduation Intelligent Manufacturing System for Advanced Planning and Scheduling in PI-enabled Hyperconnected Fixed-Position Assembly Islands

Mingxing Li¹, Daqiang Guo^{1,2}, Ray Zhong¹, G.Q. Huang¹

¹Department of Industrial and Manufacturing Systems Engineering, HKU-ZIRI Lab for Physical Internet, The University of Hong Kong, Hong Kong, China

²Department of Mechanical and Energy Engineering, Southern University of Science and Technology, Shenzhen, China

Corresponding author: gqhuang@hku.hk

Abstract: *The inherent complexity and uncertainty of planning and scheduling problems in production management have plagued researchers and practitioners for decades, especially when confronted with more diversified customer demand, fast-changing supply chain and market. Physical Internet (PI) enabled manufacturing shows the potential to revolutionize the way production and operations are managed and done in factories. Massive production data and information are real-time accessible for decision-makers thanks to the application of various frontier sensing, networking, and computing technologies in PI. Thus, it is imperative to study how to leverage the strengths of real-time data and information to support production planning and scheduling in PI-enabled manufacturing environment. This paper proposes a five-phase framework of the Graduation Intelligent Manufacturing System (GiMS) to facilitate decision-making in the PI-enabled fixed-position assembly islands. GiMS divides space and time scopes of a factory into finite areas and intervals to minimize complexity and approximate uncertainty, so that the original monolithic planning and scheduling decision can be discretized into a series of real-time decisions. Shop floor status is established and updated in real-time through PI-enabled visibility, traceability, and hyperconnectivity. Synchronization mechanisms of GiMS are designed to provide globally optimized and locally resilient solutions. Finally, a numerical study is carried out. The results show that GiMS has a well-balanced and stable performance on several measures in a dynamic environment.*

Keywords: *Graduation Intelligent Manufacturing System (GiMS), Advanced planning and scheduling, PI-enabled manufacturing, Synchronization*

1 Introduction

The layout of fixed-position assembly islands (FPAI) is commonly found in fragile and bulky equipment production such as aircraft, rotary printing presses, and lifts (Guo et al., 2020b). The FPAI consists of several assembly islands. Operators move from one island to another to perform specific operations. The product usually remains at one island for its entire assembly process to reduce costs and avoid potential damage in the movement, while the required components and equipment are moved to the island according to the assembly plan. Thus, there is no need to consider the traditional assembly line balancing problem in FPAI (Zeltzer et al., 2017). However, the increasing customized demand, inappropriate assembly islands configuration, frequent setups, and long waiting times complicate the production and operations management (planning, scheduling, execution, and control) in FPAI. Moreover, sophisticated

assembly operations, multiple production resources movements (e.g., parts, tools) are error-prone so that the stochastic nature is further amplified.

Considerable attention has been devoted to addressing the production complexity and uncertainty from both industry and academia. Leading manufacturers and practitioners have developed a wide variety of advanced manufacturing systems such as enterprise resource planning (ERP) systems and manufacturing execution systems (MES) to generate production plans and schedules (Stratman, 2007). The usefulness of these systems is widely appreciated, but they are gradually becoming insufficient to meet the current customer demand with mixed production volumes and increasing product variety. Moreover, it is hard for using these systems to respond to the disturbances and uncertain issues in actual production progress without utilizing real-time shop floor data promptly (Land, 2009). Besides, frequent rescheduling may cause resistance to change, which might be counterproductive in improving production efficiency (Rahmani & Ramezani, 2016). Then again, some manufacturers invested massively to build highly automated production lines. Still, the performance fails to come up to expectations because automation is perfect for executing static schedules with its preciseness and efficiency. However, automation alone is not smart enough to deal with various uncertainties in a nowadays dynamic and stochastic production environment; effective coordination and synchronization are indispensable (Yang et al., 2019).

On the other hand, researchers have tried a large variety of methods and algorithms to solve the APS problems mathematically and computationally. These approaches have produced more or less similar results that were theoretically optimal or near-optimal, but their performance in practice is unstable because the uncertainties are not well tackled. Afterward, hierarchical and production planning and scheduling (HPPS) and multi-period production planning and scheduling with rolling horizons (MPRH) have emerged. HPPS and MPRH are two typical "divide and conquer" approaches to locate and manage complexity and uncertainty. HPPS decomposes a complex problem into small size subproblems while MPRH discretizes the planning horizon into multiple short time intervals with manageable uncertainty (Campbell, 1992; Hax & Meal, 1973). However, both MPRH and HPPS require but suffer from the lack of feedback and updating mechanisms to integrate subproblems (McKay et al., 1995; Omar & Teo, 2007).

Fortunately, the Physical Internet (PI) enabled manufacturing shows tremendous potential in revolutionizing the way production and operations are managed and done in factories (Lin et al., 2018a). PI was first mentioned in the domain of logistics (Markillie, 2006) and has now been widely used for transforming logistics management worldwide. Recently, the concept of PI is extended to the manufacturing shop floor where cutting-edge technologies such as Industrial Internet-of-Things (IIoT), Cloud Computing (CC), Digital Twin (DT) are deployed (Zhong et al., 2017a). These disruptive technologies promise to upgrade the traditional manufacturing objects to smart objects augmented with identification, sensing, and network capabilities (Luo et al., 2019b; Zhong et al., 2017b). Real-time data and information are accessible. The connectivity, visibility, and traceability of the PI-enabled manufacturing environment bring new hope to break the bottleneck of complexity and uncertainty in production and operations management. To achieve real-time advanced planning and scheduling (APS) in FPAI, there are several research challenges that need to be resolved.

Firstly, how the APS problem is reshaped the PI-enabled manufacturing environment? How to resolve the complex and stochastic nature of production and operation management by capitalizing on the revolutionary power of real-time information and data in the PI-enabled FPAI?

Secondly, how to develop a general framework and solution for manufacturing optimization problems considering their new features in the PI-enabled hyperconnected FPAI? How to discretize the traditional monolithic APS decision into a series of real-time decisions, and how to establish their connections and dependencies using real-time visibility and traceability?

Thirdly, how to design a real-time job allocation and execution mechanism considering the actual shop floor situation such as the availability of men, machines, materials to organize production activities in a smooth and resilient manner in FPAI?

This paper proposes a five-phase framework of the Graduation Intelligent Manufacturing System (GiMS) for achieving Real-Time Advanced Planning and Scheduling (RT-APS) in PI-enabled hyperconnected FPAI. GiMS divides the space and time scopes of a factory into finite areas and intervals to localize disturbances and approximate uncertainties so that the original complex optimization problem is discretized to a series of subproblems with different spatiotemporal characteristics. Corresponding synchronization mechanisms of GiMS are designed to generate a global solution and support real-time decision making at both managerial and operational level. Finally, a numerical study is conducted to evaluate the performance of the proposed method. And several managerial insights are offered based on the results.

The rest of this article is organized as follows. Section 2 reviews relevant literature. The GiMS-enabled FPAI are presented in section 3. Section 4 explains the detailed synchronization mechanisms of GiMS. The numerical study is conducted in section 5. Finally, section 6 summarizes the paper and gives future perspectives.

2 Literature Review

2.1 IIoT-enabled Smart Manufacturing

The concept of IIoT is closely related to PI-enabled manufacturing (Zhong et al., 2017a). As one of the cutting-edge technologies in industry 4.0, IIoT promises to construct seamless connectivity between production elements and improve data and information visibility of the shop floor (Guo et al., 2020c). The pioneering researches on IIoT application in manufacturing include identification technologies such as Auto-ID, RFID. Udoka (1991) presented an overview of automated data capture technologies and claimed that these technologies are critical to the success of automated manufacturing systems. Huang et al. (2008) proposed a RFID-enabled wireless manufacturing framework to improve operational efficiency in adaptive assembly planning and control. Zhong et al. (2013) designed a real-time RFID-enabled MES to support real-time production decisions. Lin et al. (2018b) applied iBeacon technologies that provided real-time visibility to facilitate decision-making for supervisors and daily operations for workers. More recently, the potential of industrial wearables is also widely investigated (Kong et al., 2019; Li et al., 2019). Thanks to the extensive applications of IIoT technologies, vast amounts of production data are accessible. Kusiak (2017) and Kuo and Kusiak (2019) revealed the importance of big data in smart manufacturing and identified five gaps to filled for realizing the next industrial revolution.

Indeed, manufacturing is getting smarter with the deployment of IIoT devices, and massive production data are real-timely available in such an environment. However, IIoT is not readily served as an effective mechanism for the conversion from data to valuable information and knowledge. Besides, how IIoT technologies reshape the APS problems and how real-time data can facilitate APS decision considering the new features of these problems in a PI-enabled manufacturing environment need further researches.

2.2 Hierarchical/Multi-Period Planning and Scheduling

It is acknowledged that the APS problems are complex and stochastic (Efthymiou et al., 2016; Keller & Bayraksan, 2009). Researchers realize that the breakthrough to next-generation manufacturing is impossible without overcoming the bottleneck of complexity and uncertainty. HPPS and MPRH are two typical approaches to manage complexity and uncertainty.

The core idea of HPPS is to decompose planning and scheduling problem into subproblems with limited complexity and uncertainty (Bitran et al., 1982; Dempster et al., 1981). Omar and Teo (2007) developed a three-level hierarchical approach in a batch process environment, and the proposed method is tested and validated using industrial data. More recently, O'Reilly et al. (2015) presented an overall decision-making framework for small- and medium-sized food manufacturers with the applications of HPPS. Menezes et al. (2016) studied planning and scheduling problem in bulk cargo terminals and proposed a hierarchical approach with a mathematical model for integration. MPRH discretizes the planning and scheduling horizon into multiple time periods which are short enough with manageable uncertainty (Sridharan et al., 1987). Balakrishnan and Cheng (2007) gave a comprehensive review of research to address the reconfiguration and uncertainty issues in cellular manufacturing under conditions of multi-period planning horizons. Torkaman et al. (2017) presented MIP-based heuristic models with rolling horizons to solve a multi-stage multi-product multi-period capacitated flow shop planning problem with lot sizing.

HPPS and MPRH were proven to be effective in traditional manufacturing settings. Nevertheless, the two methods are hampered by the amplified complexity and uncertainty in the modern manufacturing environment because both required but suffered from the lack of updating mechanisms to integrate subproblems (Omar & Teo, 2007).

2.3 Manufacturing Synchronization

The emergence of manufacturing synchronization (MfgSync) provides a new perspective of production and operations management in the era of industry 4.0. The idea of MfgSync is mostly related to just-in-time (JIT) philosophy, which was firstly proposed by Toyota (Sugimori et al., 1977). JIT advocates that "all processes produce the necessary parts at the necessary time and have on hand only the minimum stock necessary to hold the processes together." JIT aims to reduce the production lead time and inventory level. In contrast, MfgSync focuses on synchronized order-job-operation management to achieve an overall-well performance regarding cost-efficiency, simultaneity, and punctuality (Guo et al., 2020a).

Relevant research on MfgSync is relatively recent. Riezebos (2011) examined the effectiveness of some new heuristics that are based on insights from assembly system design and workload control, and compare their performance with an optimal solution approach. Chen et al. (2019b) studied MfgSync of scheduling dynamic arrival orders in a hybrid flow shop. Luo et al. (2019a) investigated synchronized scheduling problem of make to order plant and cross-docking warehouse and provided decision-maker with managerial insights to configure the production resource and warehousing resource in different scenarios. In addition, production-logistics, production-shipment synchronization also attracted widespread research attention (Chen et al., 2019a; Luo et al., 2017; Qu et al., 2016).

The concept of MfgSync inspires this study to incorporate real-time synchronization mechanisms into the framework of GiMS to facilitate decision-making in planning, scheduling, execution, and control for PI-enabled hyperconnected FPAI.

3 Graduation Intelligent Manufacturing System for Physical Internet-enabled Hyperconnected Fixed Position Assembly Islands

In this section, the five-phase framework of GiMS for PI-enabled hyperconnected fixed-position assembly islands is developed. The configuration and layout of FPAI are introduced in section 3.1. Section 3.2 illustrates the graduation manufacturing systems for FPAI. Lastly, the five-phase framework to implement GiMS in FPAI is given in section 3.3.

3.1 Typical Layout and Configuration of FPAI

FPAI is a typical manufacturing mode for producing bulky or fragile items. As shown in Figure 1, there are two main areas in FPAI. One area (section 1 in the figure) contains the main manufacturing resources, including operators, equipment, and materials. The other area (section 2 in the figure) illustrates the whole assembly process at a single island. Due to the space limitation, each assembly island is only able to keep the work-in-process, a toolbox, a buffer for one part and consumable materials needed for the assembly task, and some space for the operator. The FPAI requires frequent and accurate movement of manufacturing resources, and the assembly task can be started only when all the required materials, tools and operators are ready.

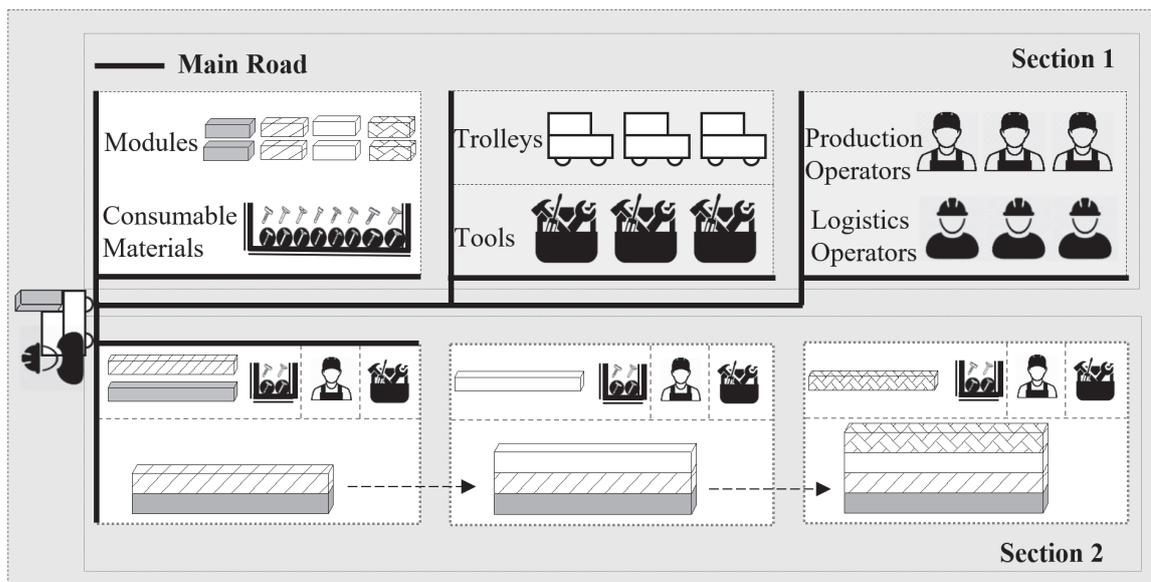


Figure 1: The Layout and Configuration of Fixed-Position Assembly Islands

3.2 Graduation Manufacturing System for FPAI

Graduation Manufacturing System is a novel manufacturing mode that is inspired by the graduation ceremony. Its basic form and principles can be found in previous research (Guo et al., 2020; Lin, et al., 2018). The GMS mode is applied in FPAI.

By analogy to three kinds of tickets used in the graduation ceremony, three kinds of tickets including job ticket (JT), setup ticket (ST), operation ticket (OT) and twined logistics ticket (LT) correspond to admission tickets, program tickets and name tickets are designed in GMS. Figure 2 illustrates the procedure. The workshop production activities are organized and managed through JTs, STs, OTs and LTs with simplicity and resilience. JTs are generated based on real-time customer demand and production constraints to achieve dynamic workload control.

Flexible control of setup can be achieved using STs to avoid unnecessary waiting time and unreasonable setup, and the setup operation can be informed in advance and performed at the right time. OTs and twined LTs ensure the synchronization and coordination of production and logistics processes for JIT delivery.

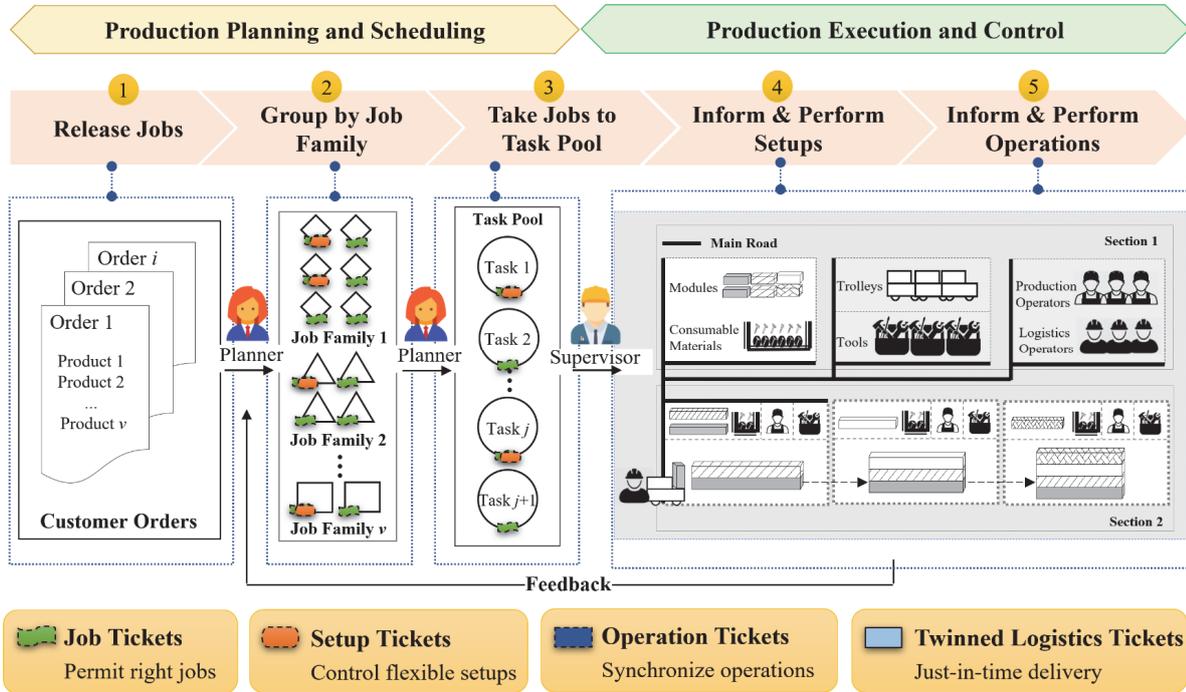


Figure 2: GMS for Fixed-Position Assembly Islands

3.3 The Five-Phase Framework of GiMS for FPAI

Unlike the simple and visible environment, repetitive operation in the graduation ceremony, the actual shop floors are far more complex and dynamic, and the production processes are sophisticated. Therefore, to fully tap the potential of GMS, the real-time information visibility and traceability of the shop floor, real-time coordination and communication among different parties, and real-time synchronization mechanisms for decision-making are indispensable. To overcome these challenges, as shown in Figure 3, this section presents the GiMS with five key phases, namely, Finite meshing, Smart digitization, Out-of-Order ticketing, Visibility and traceability analytics, and Synchronization, for achieving real-time planning and scheduling in PI-enabled FPAI.

The finite meshing phase is to divide the organization and decision horizon of the FPAI workshop into several graduation ceremony stages (GCS, a number of assembly islands in a short time period). Each GCS is composed of several assembly islands. The overall planning horizon is discretized into shorter scheduling periods, which is several hours. This phase minimizes the complexity and localizes uncertainties through spatiotemporal discretization. Operation elements (men, machines, materials) are defined for all "GCS" as physical twins.

The second phase digitizes the operation elements at all GCS for generating digital twins with the PI technologies. With the deployment of mobile crowdsensing and IIoT devices, all physical entities are digitized to provide the FPAI with greater interconnectedness of resources, better circulation of production information flow, and data flow as a solid foundation for a higher level of synchronization. The PI-FPAI is characterized by hyperconnectivity between physical entities, high-quality real-time data, and information acquisition.

The Out-of-Order (OoO) ticketing phase implements the processing logic of operation elements. The OoO in factories organizes the onsite production execution in an order governed by the availability of materials, machines, and men. Operators look ahead in a window of jobs through smart devices and find those that are ready to be processed. The key features of OoO are the high degree of autonomy, flexibility, and resilience at the operational level. The adverse effects of uncertainties can be minimized under OoO ticketing, which will be further discussed in section 4.

The cyber-physical visibility and traceability (CPVT) analytics is utilized to identify and establish the dependencies and connectivity of GCS and operation elements and to mitigate the spatiotemporal uncertainties. The dependency and connectivity usually refer to the logical relationship between GCS, such as how the state of ticket pools update over time and how the tickets flow between elements. These real-time data and information are vital for supporting synchronized decision-making.

The last phase designs the synchronization mechanisms under GiMS to facilitate upper-level planning and scheduling (synchronized ticket allocation) and lower-level onsite execution and control (real-time ticket sequencing). Detailed models and algorithms will be presented in Section 4. The bi-level synchronization mechanism promises both optimized decisions at the managerial level and resilience execution, flexible control at the operational level.

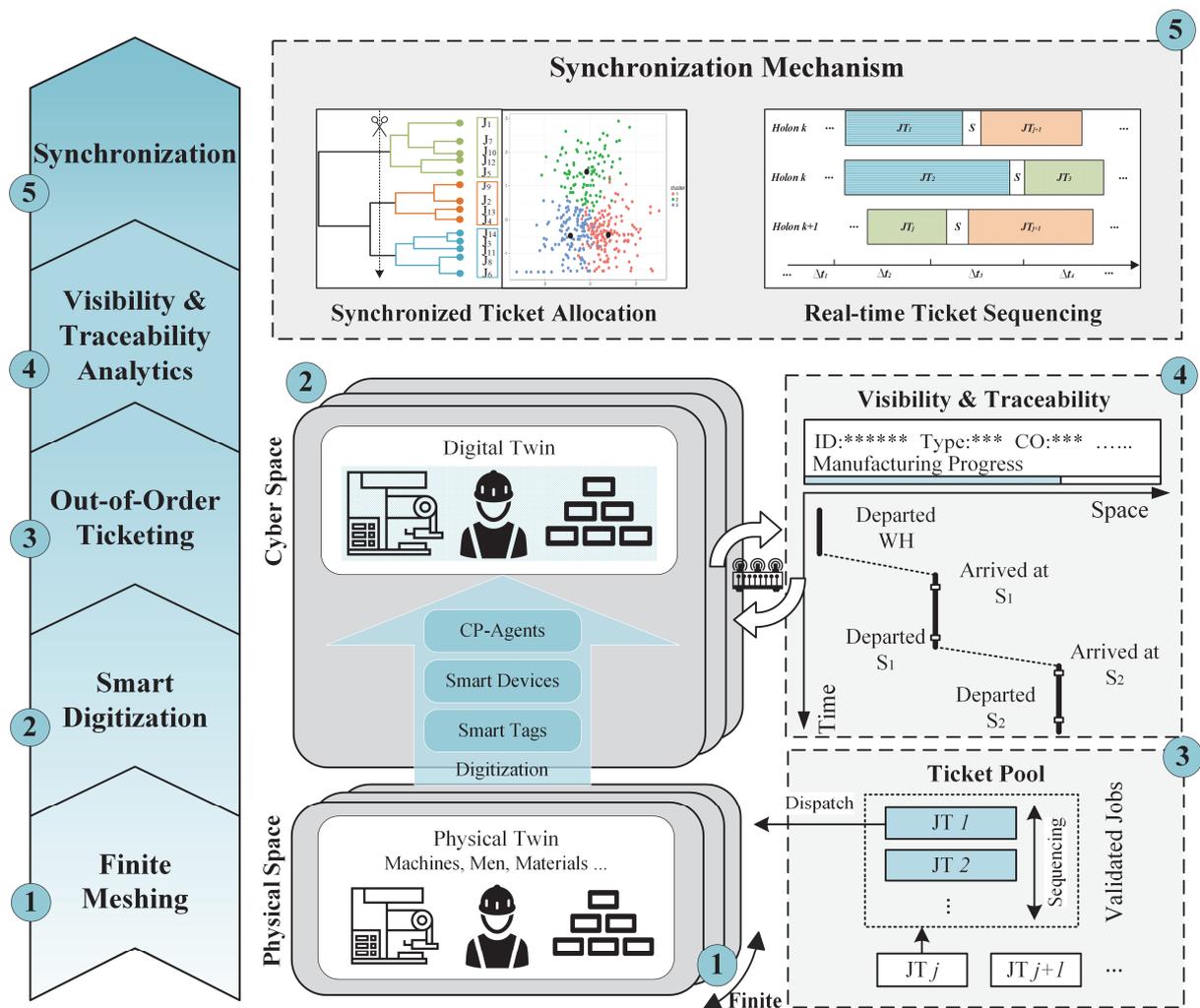


Figure 3: Five-Phase Framework of GiMS

4 Synchronization Mechanisms of GiMS

The synchronization mechanisms and algorithms are the core of GiMS. The original intention of GiMS is to cope with shifting events by sticking to a fundamental principle. Instead of generating a rigid production plan and schedule, similar jobs are clustered into job families and assigned to each GCS in GiMS. It is precisely because the jobs in the same cluster are similar, an exact processing sequence within the cluster is less significant. In section 4.1, the synchronized ticket allocation mechanism is designed for planning and scheduling at the managerial decision level. In section 4.2, the real-time ticket sequencing is developed for onsite execution and control at the operational decision level.

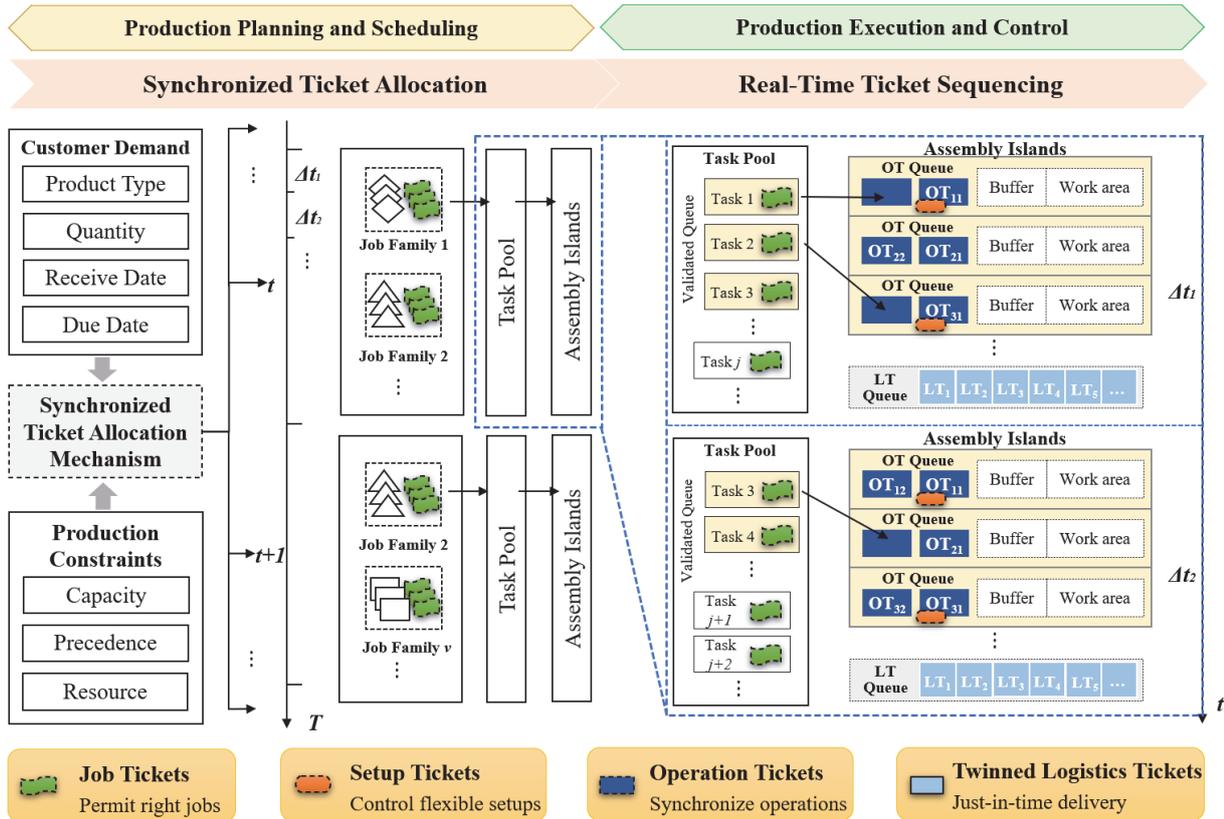


Figure 4: Bi-level Synchronization Mechanisms under GiMS

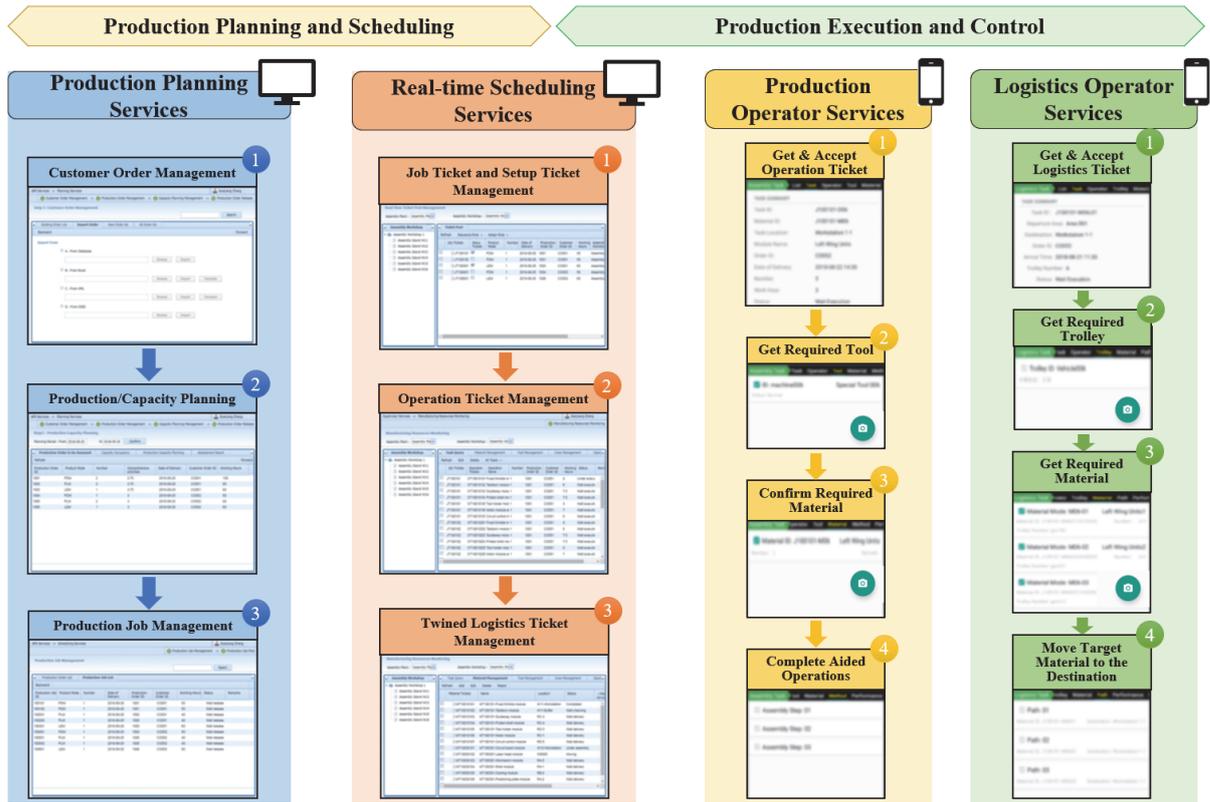


Figure 5: DApp and MApp of GiMS

4.1 Synchronized Ticket Allocation under GiMS

The left half of Figure 4 depicts the production planning and scheduling phase under GiMS. The overall planning horizon T is discretized into multiple shorter scheduling periods t to reduce complexity, localize, and approximate the uncertainties from customers, suppliers, and the market in the time dimension. Based on the real-time traceability and visibility of customer demand (e.g., product type, quantity, receive date and due date) and production constraints (e.g., capacity, resource, and precedence constraints) in period t , the synchronized ticket allocation mechanism aims to generate schedule for period $t + 1$ on an aggregate basis for families of jobs (identical or similar products) and allocate job tickets. The similarity among jobs is usually measured from the aspects of setup, material requirement, operator skill requirement, due date, correlative orders and so on (Lin et al., 2018b). Methods such as clustering analysis, mathematical programming, auction-based approaches are commonly applied at this stage. Since the scheduling period t is short relative to the planning horizon T and the exact processing sequence is not considered at this moment, the less computational effort is required. Thus, the ticket allocation mechanism can be run in near real-time to provide shop floor supervisors and managers with advanced planning and scheduling services for coping with frequent changes from customers, suppliers, and the market promptly, reliably, and positively. Besides, this mechanism allows great flexibility for job execution and progress control at the operational level.

In the FPAI case, the jobs within the same customer order and the jobs that require less setup time for changeover tend to be clustered. The hierarchical clustering is adopted, the distances of jobs are given as:

$$d(i, j) = w_1 \cdot d_1 + w_2 \cdot d_2 \quad (1)$$

Where d_1 takes value 0 if the two jobs are in the same order, 1 otherwise. d_2 is positively correlated with the setup time for changeover between job i, j . w_1 and w_2 are the weighting factors of d_1 and d_2 respectively.

And then, linkages are generated between pairs of jobs that are close together to form binary job clusters. These newly formed binary job clusters are further linked to each other to create bigger clusters until all the jobs are linked together to form a hierarchical tree. And the similarity of clusters a, b is given as

$$d(a, b) = \sqrt{\frac{2n_a n_b}{n_a + n_b}} \|\bar{a} - \bar{b}\|_2 \quad (2)$$

Where the \bar{a} and \bar{b} denote the centroids of clusters a and b , n_a and n_b the number of jobs in clusters a and b .

The synchronized ticket allocation mechanism is integrated in GiMS as the production planning services and real-time scheduling services that can be accessed through Desktop Application (DApp) to support managerial decision-making. As shown in the left half of Figure 5, The production planning service is used by planners to make short-term production plans. Three sub services, namely customer orders management, production/capacity planning, and production job management, are included. The customer orders are imported and updated through customer order management explorer, and the sequence of orders is determined by priorities (due date, the importance of the customer, etc.). The production/capacity planning service calculates which customer orders can be completed based on the available capacity within the given period. The result is converted into the production schedules and similar jobs are clustered as families that will be released into real-time task pools one by one in production job management service. The workshop production activities are organized and managed through JTs, STs, OTs, and LTs in real-time scheduling services for supervisors. Based on the real-timely monitored workload and capacity of the assembly islands, job families are released automatically and dynamically to balance the whole workshop. Once a job family is released, the JTs and STs are generated and managed in JTs and STs management service. Flexible adjustments for coping with disturbances can be made automatically or by dragging and dropping. Correspond logistics tasks, production tasks and LTs, OTs will be generated and managed in the LTs and OTs management service respectively. The systematic implementation and integration of planning and scheduling services in DApp of GiMS ensure simple but effective managerial production decision-making.

4.2 Real-Time Ticket Sequencing under GiMS

The right half of Figure 4 presents the production execution and control phase under GiMS, as the jobs allocated to a single scheduling period are similar, rigid and exact sequencing of these jobs is less significant. Thus, OoO execution of tickets is adopted to eliminate uncertainties and control the onsite production process with robustness and resilience. At the beginning of period t , correspond JTs are released to the shop floor task pool, the OTs and twined LTs for this job are activated. A JT enters the validated queue once all required resources (e.g., operator, material, machine, or tool) are available. That is, validated jobs are ready for processing. Real-time sequencing mechanism is applied to achieve synchronization at the operational level, the priority of validated tickets is computed based on Horizontal Synchronization (HSync), Vertical Synchronization (VSync) (Lin et al., 2018b), is calculated as follows:

$$P_j = w_3 * H_{sync} + w_4 * V_{sync} \quad (3)$$

Where H_{sync} represents the production progress of the order that contains job j . V_{sync} reflects the matching degree of the job j and the available assembly island k (whether the setup is required). w_3 and w_4 are the weighting factors.

Let o denote the order that contains job j , UJ_o , RJ_o , and FJ_o denote the total processing time of currently unreleased jobs, released jobs, and finished jobs of order o respectively, ε is an arbitrary small real number in case RJ_o and FJ_o equal to 0. Where f_k denotes the setup condition of assembly island k left by the previous job. H_{sync} and V_{sync} can be calculated as follows.

$$H_{sync} = UJ_o^{\frac{1}{RJ_o + FJ_o + \varepsilon}} \quad (4)$$

$$V_{sync} = S_{f_j, f_k} \quad (5)$$

When there is a vacancy in the assembly island ticket queue, the sequencing is triggered and the priority is calculated based on real-time data, the job ticket with the highest priority will be dispatched to fill the vacancy, and ST will be issued accordingly if it is required. In general, the length of the island ticket queue is set as Δt , and a ticket is "frozen" once entering the assembly island ticket queue, which means it will not be re-dispatched to avoid confusion in the shop floor. OoO guarantees smooth execution of jobs and flexible control of production progress to enhance the overall scheduling resilience in the highly dynamic and stochastic shop floor.

The real-time ticket sequencing mechanism is integrated into GiMS as the production/logistics operator services that can be accessed through Mobile Application (MApp) to facilitate onsite production execution and control. As shown in the right half of Figure 5, with the support of the real-time task pool management and production operator service, the production operator can explore the detailed OTs of the assigned JTs via the MApp on the smartphone. Once a vacancy is detected in the assembly island buffer, the validated job ticket with the highest priority will be assigned. Only the right materials are checked and confirmed, the production operator can start the specific operations with the required tools. Moreover, when suffering uncertain events, the production operator could identify and report them timely to the supervisor for further decision-making, which can minimize the impact of uncertainties. In the logistics operator service of MApp, the real-time identification and location information are available, the logistics operators can get logistics task in MApp and easily find the target trolleys and materials and move them to the designated areas. One logistics task can be submitted by the logistics operator only when the target materials are detected and confirmed at the right places at the right time through the system..

The combination of ticket allocation mechanism and real-time ticket sequencing serves as the theoretical and logical basis of GiMS, cloud services-based system implementation (DApp and MApp) guarantees the usability, adaptability, and stability of GiMS in the real-life industry with simple and effective production planning and scheduling as well as smooth and resilient onsite production execution and control. Table 1 presents the pseudocode of proposed synchronization mechanisms.

Table 1: Pseudocode of the synchronization mechanisms

Pseudocode of the synchronization mechanisms	
	Run at the beginning of each t
	Inputs: the number of GCS, the number of islands per GCS, scheduling period t , customer orders, availability of production resources
	Outputs: the jobs allocated to this period, the processing sequence of allocated jobs
1	Update the system status (order pool, job pool, finished jobs, availability of resources)
2	Estimate the production capacity of each GCS
3	Calculate the similarities among all pending jobs
4	Cluster the jobs by their similarities using Hierarchical Clustering
5	For $k \leftarrow 1$ to the number of GCS
6	Allocate clustered jobs to GCS k based on its capacity
7	Generate corresponding Job Tickets for GCS k
8	End For
9	Update unreleased jobs, released jobs to task pools
10	While the task pools are non-empty
11	Check the first available GCS k'
12	Check the first available island of GCS k'
13	For $i \leftarrow 1$ to the number of Job Tickets for GCS k'
14	Calculate the Hsync index and Vsync index
15	Update the priority of job ticket i
16	End For
17	Dispatch the job ticket with the highest priority to the first available island of GCS k'
18	Update task pools, finished jobs
19	If current time \geq next t
20	Break
21	End If
22	End While

5 Numerical Study

This section carries out a numerical study to verify the effectiveness of the proposed GiMS. Three performance measures are used in the numerical study: 1) Makespan (MS); 2) Total Setup Time (TST); 3) Total Tardiness (TTD). The experimental data are randomly generated from Table 2. The assumptions are listed as follows:

- Customer orders arrive dynamically;
- Setup is required for changeover between different product families;
- Assembly islands are homogeneous;
- Each island can only process one job at a time;
- Preemption of jobs is not allowed;
- The transportation time for jobs is negligible.

Table 2: Experimental data

Data	Value
Total number of islands	12
Number of islands per GC stage	2, 3, 4
Total number of customer orders	15, 30, 45
Number of jobs per order	20
Number of product families	10
Number of jobs left from last shift	100
Processing time (min)	U [40, 60]
Setup time (min)	U [5, 25]
Orders inter-arrival time (min)	Pois(60)
Due dates of orders	U [3, 5]×8×60

5.1 Parameters Setting

The number of islands per GCS, t , and combination of weights are important parameters of the proposed synchronization mechanisms that will be properly set through the following experiments.

The number of islands per GCS is set as 2, 3, and 4, and the t is set as 60, 120, 180 and 240 mins (see Table 3). Figure 6 gives the curves of the three measures. Generally, the tendencies of MS, TST are similar despite the number of islands per GCS. As the t extends, the MS and TST decrease, and the amplitude of variation is relatively flat when t is equal or greater than 120 mins. This is possible because there are more jobs in the pool with the dynamic arrival of orders, the possibility to find similar jobs is higher (less setup time for changeover), and more orders are considered in one period (longer waiting time for each order). It implies that it is preferable to set a larger t when the company has high setup cost. TTD shows fluctuations as t ranges from 60 to 240 mins, and lower values can be obtained when t is set as 120 and 180 mins. The number of islands per GCS has a relatively limited impact on the performance. Generally, more islands per GCS performs better on MS and TST, less islands per GCS performs better on TTD. In a comprehensive perspective, it is preferred to set 3 islands per GCS and $t = 120$ mins for the following experiments.

Table 3: The performance of the different number of islands per GCS and t

	t			
	60	120	180	240
2 islands per GCS				
MS	3242	3105	2906	2856
TST	7136	5348	3105	2637
TTD	15	0	259	1
3 islands per GCS				
MS	3270	2980	2919	2897
TST	7199	4023	3367	3070
TTD	0	299	273	383
4 islands per GCS				
MS	3196	3017	2895	2902
TST	6298	4165	3082	2849
TTD	1521	434	335	2056

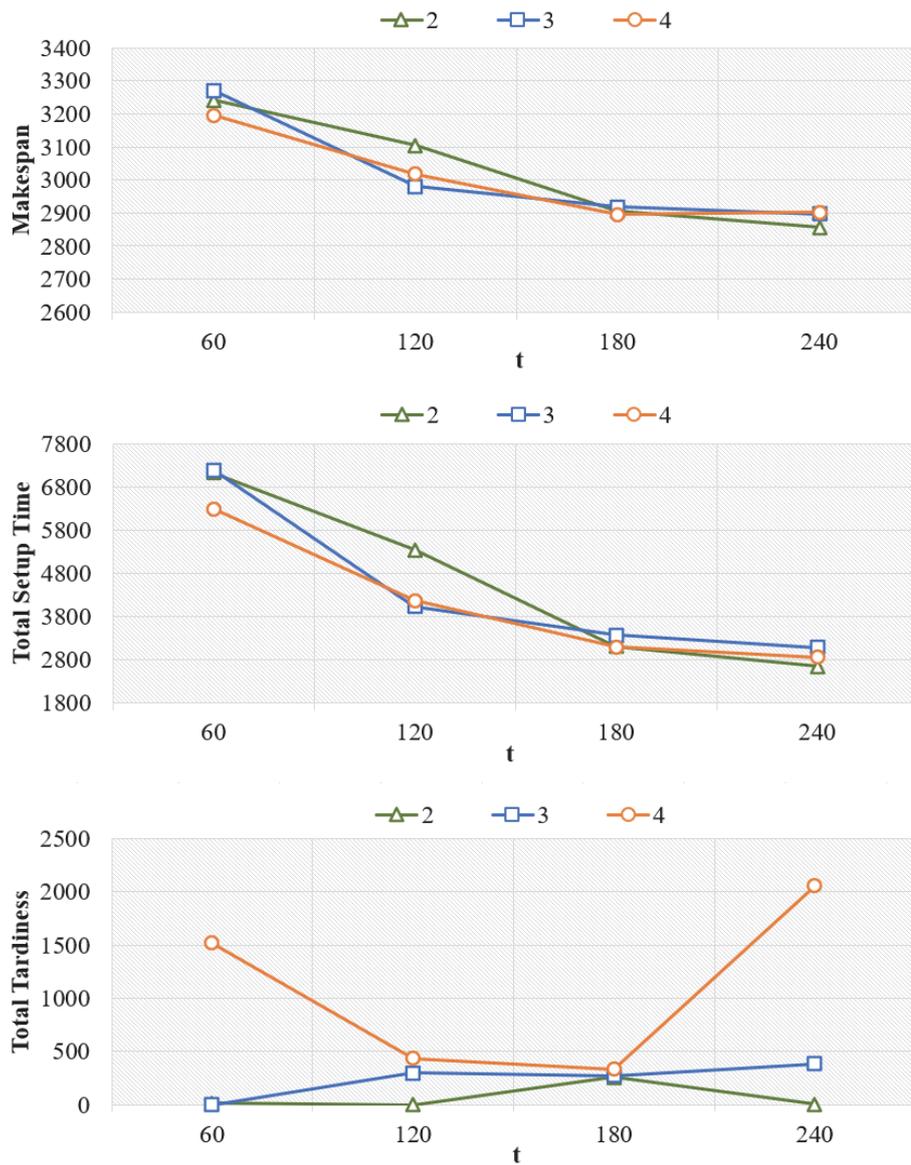


Figure 6: The performance of the different number of islands per GCS and t

There are 4 crucial weights in the proposed solution algorithm, namely (w_1, w_2) in the ticket allocation mechanism and (w_3, w_4) in real-time sequencing. In this experiment, (w_1, w_2) and (w_3, w_4) are set as $(0.25, 0.75)$, $(0.5, 0.5)$, and $(0.75, 0.25)$ respectively, in total 9 combinations of weights are considered as shown in Table 4 and Figure 7. It is found that (w_1, w_2) have a greater impact on the performance than (w_3, w_4) , this indicates that the exact processing sequence of jobs within a single period t is less significant because the jobs allocated to one period are similar, and it also means more flexibility for onsite operators to make adjustments to cope with uncertainties to achieve smooth execution and resilient control of production when (w_1, w_2) are properly set.

Table 4: The performance of different combinations of weights

w_1, w_2		w_3, w_4		
		0.25, 0.75	0.5, 0.5	0.75, 0.25
0.25, 0.75	MS	2940	2919	2931
	TST	3333	3287	3315
	TTD	169	0	186
0.5, 0.5	MS	2984	2980	2972
	TST	4010	4023	4000
	TTD	299	299	663
0.75, 0.25	MS	3073	3047	3070
	TST	5046	4834	4837
	TTD	0	0	0

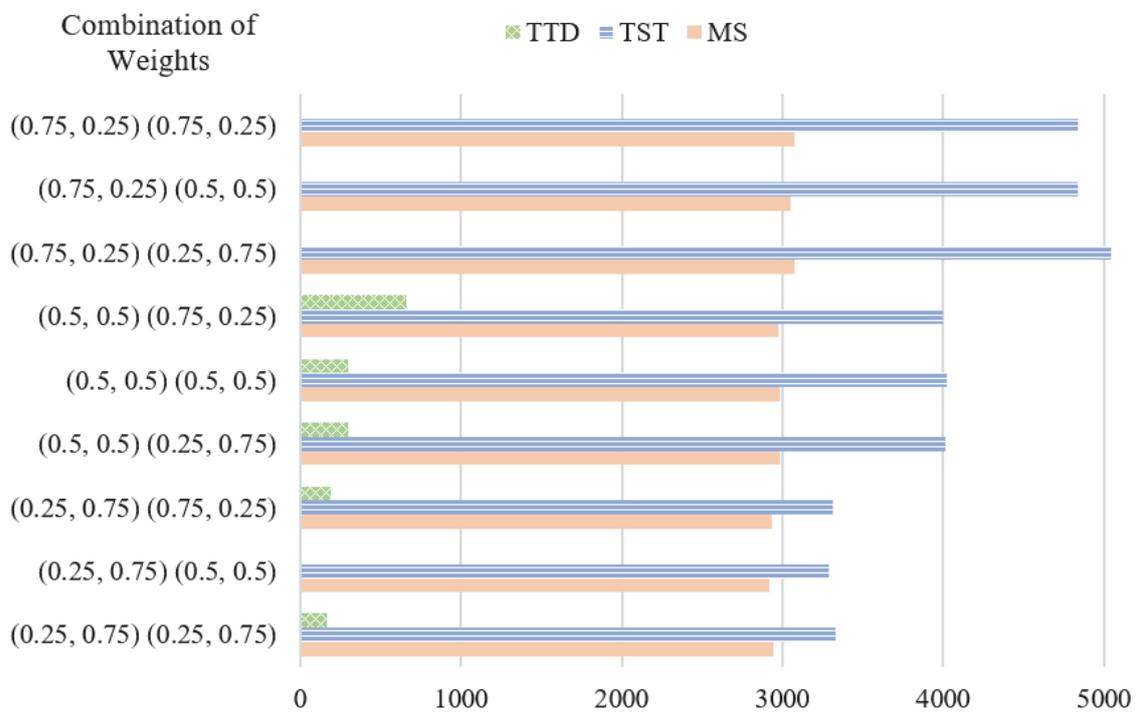


Figure 7. The performance of different combinations of weights

5.2 Performance Evaluation

The experiment below evaluates the performance of the proposed method under various configurations. Three typical dispatching rules are adopted as references.: 1) Earliest Due Date (EDD), which is found to be effective in reducing TTD; 2) Shortest Processing Time (SPT), which is one of most classical rules in literature; 3) First-Come First-Served (FCFS), customer orders are processed one by one under FCFS. One important reason for choosing these rules is that they require less computation efforts. In real-time decision-making, the time-limit must be considered (Ghaleb et al., 2020). The proposed method can generate results in a few seconds (even for large size instances), which means it can react to disturbances promptly.

Table 5 presents the results. It is observed that the proposed GiMS outperforms other dispatching rules in terms of MS, TST, TTD, and backlog jobs. The EDD rule suppose to have better performance on TTD and backlog jobs as the most urgent jobs are processed first under this rule. It might be attributed to the frequent setups that actually prolong the whole production

process (the TST under EDD is more than triple that under GiMS, and the MS under EED is about 20% longer than that under GiMS). This experiment indicates that the performance of GiMS is well-balanced and relatively stable in a dynamic environment. Besides, the advantages of GiMS are more obvious in normal and high demand.

Table 5: The performance evaluation of GiMS

	GiMS	EDD	SPT	FCFS
15 Customer Orders (Low demand)				
MS	1514	1833	1687	1724
TST	2151	6066	4448	4876
TTD	0	0	81	0
Backlog Jobs	0	0	2	0
30 Customer Orders (Normal demand)				
MS	2919	3567	3472	3442
TST	3287	11215	10088	9692
TTD	0	1705	92567	22388
Backlog Jobs	0	27	178	84
45 Customer Orders (High demand)				
MS	4497	5273	5302	5141
TST	6444	15931	16269	14397
TTD	18709	133279	539406	148962
Backlog Jobs	88	368	439	304

5.3 Managerial Insights

Based on the numerical results, several managerial insights can be concluded for practitioners. Firstly, the real-time manufacturing data and information provide the managers and supervisors with real-time visibility and traceability. The GiMS is proved effective to generate real-time APS decisions using visibility and traceability. Secondly, in comparison with traditional planning and scheduling strategies, GiMS can obtain overall balanced and stable solutions regarding multiple measures in a dynamic environment. The advantages of GiMS are more obvious in normal and high demand. Thirdly, key parameters should be carefully adjusted according to the actual conditions of the factory. The schedule will be too rigid to lose resilience when space and time scope of a GCS is too small; while it is lack responsiveness to frequent disturbances and the similarity of clustered jobs is weakened if the size is too large.

6 Conclusion

In conclusion, this study has investigated the advanced planning and scheduling problem in the PI-enabled fixed-position assembly islands. With the deployment of IIoT devices, the PI-enabled manufacturing environment is characterized by great visibility, traceability, and hyperconnectivity. A five-phase framework of GiMS with synchronization mechanisms is proposed to tackle the complexity and uncertainty of production and operations management in modern manufacturing. By minimizing the complexity and uncertainty in meshing, the original monolithic APS decision can be discretized into a series of real-time decisions. A global solution is obtained through synchronization mechanisms. Finally, a numerical study was carried out to examine the superiority of GiMS. The results showed that GiMS had a well-balanced and stable performance on selected measures in a dynamic environment.

The contributions of this paper are threefold: 1) It proposed a formalized five-phase framework of GiMS, which can leverage the revolutionary power of real-time data to support production and operations management in the PI-enabled manufacturing environment; 2) It designed the

synchronization mechanisms of GiMS to facilitate real-time decision-making at both managerial and operational level; 3) It carried out a numerical study to verify the effectiveness of GiMS and offers some managerial implications based on the results.

Two possible research directions deserve further explorations. On the one hand, a more comprehensive case study might be carried out in the future to consider various uncertain events, adopt more measures like schedule stability and robustness. On the other hand, the application of the five-phase framework of GiMS may be extended to other manufacturing layouts such as job shop and flow shop. It might even be applied to solve other combinatorial optimization problems such as vehicle routing encountered in the real-world.

Acknowledgements

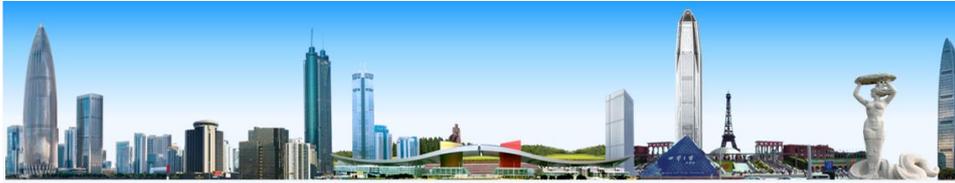
Acknowledgement to Zhejiang Provincial, Hangzhou Municipal, Lin'an City Governments, Hong Kong ITF Innovation and Technology Support Program (ITP/079/16LP) and partial financial support from the 2019 Guangdong Special Support Talent Program – Innovation and Entrepreneurship Leading Team (China) (2019BT02S593).

References

- Balakrishnan, J., & Cheng, C. H. (2007). Multi-period planning and uncertainty issues in cellular manufacturing: A review and future directions. *European Journal of Operational Research*, 177(1), 281-309.
- Bitran, G. R., Haas, E. A., & Hax, A. C. (1982). Hierarchical production planning: A two-stage system. *Operations Research*, 30(2), 232-251.
- Campbell, G. M. (1992). Master production scheduling under rolling planning horizons with fixed order intervals. *Decision Sciences*, 23(2), 312-331.
- Chen, J., Huang, G. Q., & Wang, J.-Q. (2019a). Synchronized scheduling of production and outbound shipping using bilevel-based simulated annealing algorithm. *Computers & Industrial Engineering*, 137, 106050.
- Chen, J., Wang, M., Kong, X. T., Huang, G. Q., Dai, Q., & Shi, G. (2019b). Manufacturing synchronization in a hybrid flowshop with dynamic order arrivals. *Journal of Intelligent Manufacturing*, 30(7), 2659-2668.
- Dempster, M. A. H., Fisher, M., Jansen, L., Lageweg, B., Lenstra, J. K., & Rinnooy Kan, A. (1981). Analytical evaluation of hierarchical planning systems. *Operations Research*, 29(4), 707-716.
- Efthymiou, K., Mourtzis, D., Pagoropoulos, A., Papakostas, N., & Chryssolouris, G. (2016). Manufacturing systems complexity analysis methods review. *International Journal of Computer Integrated Manufacturing*, 29(9), 1025-1044.
- Ghaleb, M., Zolfagharinia, H., & Taghipour, S. (2020). Real-time production scheduling in the Industry-4.0 context: Addressing uncertainties in job arrivals and machine breakdowns. *Computers & Operations Research*, 123, 105031.
- Guo, D., Li, M., Zhong, R., & Huang, G. (2020a). Graduation Intelligent Manufacturing System (GiMS): an Industry 4.0 paradigm for production and operations management. *Industrial Management & Data Systems*.

- Guo, D., Zhong, R. Y., Lin, P., Lyu, Z., Rong, Y., & Huang, G. Q. (2020b). Digital twin-enabled Graduation Intelligent Manufacturing System for fixed-position assembly islands. *Robotics and Computer-Integrated Manufacturing*, 63, 101917.
- Guo, D., Zhong, R. Y., Ling, S., Rong, Y., & Huang, G. Q. (2020c). A roadmap for Assembly 4.0: self-configuration of fixed-position assembly islands under Graduation Intelligent Manufacturing System. *International Journal of Production Research*, 1-16.
- Hax, A. C., & Meal, H. C. (1973). Hierarchical integration of production planning and scheduling. *Massachusetts Institute of Technology, Sloan School of Management*.
- Huang, G. Q., Zhang, Y., Chen, X., & Newman, S. T. (2008). RFID-enabled real-time wireless manufacturing for adaptive assembly planning and control. *Journal of Intelligent Manufacturing*, 19(6), 701-713.
- Keller, B., & Bayraksan, G. (2009). Scheduling jobs sharing multiple resources under uncertainty: A stochastic programming approach. *IIE Transactions*, 42(1), 16-30.
- Kong, X. T., Luo, H., Huang, G. Q., & Yang, X. (2019). Industrial wearable system: the human-centric empowering technology in Industry 4.0. *Journal of Intelligent Manufacturing*, 30(8), 2853-2869.
- Kuo, Y.-H., & Kusiak, A. (2019). From data to big data in production research: the past and future trends. *International Journal of Production Research*, 57(15-16), 4828-4853.
- Kusiak, A. (2017). Smart manufacturing must embrace big data. *Nature*, 544(7648), 23-25.
- Land, M. J. (2009). Cobacabana (control of balance by card-based navigation): A card-based system for job shop control. *International Journal of Production Economics*, 117(1), 97-103.
- Li, M., Xu, G., Lin, P., & Huang, G. Q. (2019). Cloud-based mobile gateway operation system for industrial wearables. *Robotics and Computer-Integrated Manufacturing*, 58, 43-54.
- Lin, D., Lee, C. K., Lau, H., & Yang, Y. (2018a). Strategic response to Industry 4.0: an empirical investigation on the Chinese automotive industry. *Industrial Management & Data Systems*.
- Lin, P., Shen, L., Zhao, Z., & Huang, G. Q. (2018b). Graduation manufacturing system: synchronization with IoT-enabled smart tickets. *Journal of Intelligent Manufacturing*, 1-16.
- Luo, H., Wang, K., Kong, X. T., Lu, S., & Qu, T. (2017). Synchronized production and logistics via ubiquitous computing technology. *Robotics and Computer-Integrated Manufacturing*, 45, 99-115.
- Luo, H., Yang, X., & Wang, K. (2019a). Synchronized scheduling of make to order plant and cross-docking warehouse. *Computers & Industrial Engineering*, 138, 106108.
- Luo, S., Liu, H., & Qi, E. (2019b). Big data analytics-enabled cyber-physical system: model and applications. *Industrial Management & Data Systems*.
- Markillie, P. (2006). *The physical internet: A survey of logistics*: Economist Newspaper.
- McKay, K. N., Safayeni, F. R., & Buzacott, J. A. (1995). A review of hierarchical production planning and its applicability for modern manufacturing. *Production Planning & Control*, 6(5), 384-394.

- Menezes, G. C., Mateus, G. R., & Ravetti, M. G. (2016). A hierarchical approach to solve a production planning and scheduling problem in bulk cargo terminal. *Computers & Industrial Engineering*, 97, 1-14.
- O'Reilly, S., Kumar, A., & Adam, F. (2015). The role of hierarchical production planning in food manufacturing SMEs. *International Journal of Operations & Production Management*, 35(10), 1362-1385.
- Omar, M. K., & Teo, S. (2007). Hierarchical production planning and scheduling in a multi-product, batch process environment. *International Journal of Production Research*, 45(5), 1029-1047.
- Qu, T., Lei, S., Wang, Z., Nie, D., Chen, X., & Huang, G. Q. (2016). IoT-based real-time production logistics synchronization system under smart cloud manufacturing. *The International Journal of Advanced Manufacturing Technology*, 84(1-4), 147-164.
- Rahmani, D., & Ramezani, R. (2016). A stable reactive approach in dynamic flexible flow shop scheduling with unexpected disruptions: A case study. *Computers & Industrial Engineering*, 98, 360-372.
- Riezebos, J. (2011). Order sequencing and capacity balancing in synchronous manufacturing. *International Journal of Production Research*, 49(2), 531-552.
- Sridharan, V., Berry, W. L., & Udayabhanu, V. (1987). Freezing the master production schedule under rolling planning horizons. *Management Science*, 33(9), 1137-1149.
- Stratman, J. K. (2007). Realizing benefits from enterprise resource planning: does strategic focus matter? *Production and Operations Management*, 16(2), 203-216.
- Sugimori, Y., Kusunoki, K., Cho, F., & Uchikawa, S. (1977). Toyota production system and kanban system materialization of just-in-time and respect-for-human system. *The International Journal of Production Research*, 15(6), 553-564.
- Torkaman, S., Ghomi, S. F., & Karimi, B. (2017). Multi-stage multi-product multi-period production planning with sequence-dependent setups in closed-loop supply chain. *Computers & Industrial Engineering*, 113, 602-613.
- Udoka, S. J. (1991). Automated data capture techniques: a prerequisite for effective integrated manufacturing systems. *Computers & Industrial Engineering*, 21(1-4), 217-221.
- Yang, H., Kumara, S., Bukkapatnam, S. T., & Tsung, F. (2019). The internet of things for smart manufacturing: A review. *IIE Transactions*, 1-27.
- Zeltzer, L., Aghezzaf, E.-H., & Limère, V. (2017). Workload balancing and manufacturing complexity levelling in mixed-model assembly lines. *International Journal of Production Research*, 55(10), 2829-2844.
- Zhong, R. Y., Dai, Q., Qu, T., Hu, G., & Huang, G. Q. (2013). RFID-enabled real-time manufacturing execution system for mass-customization production. *Robotics and Computer-Integrated Manufacturing*, 29(2), 283-292.
- Zhong, R. Y., Xu, C., Chen, C., & Huang, G. Q. (2017a). Big data analytics for physical internet-based intelligent manufacturing shop floors. *International Journal of Production Research*, 55(9), 2610-2621.
- Zhong, R. Y., Xu, X., Klotz, E., & Newman, S. T. (2017b). Intelligent manufacturing in the context of industry 4.0: a review. *Engineering*, 3(5), 616-630.



Design and decision optimization of a robot shuttle system

Wei Wang¹ and Yaohua Wu¹ and Ming Li²
Shandong University, Jinan, China
Shandong Jianzhu University, Jinan, China
Yaohua Wu: ww_sdu@mail.sdu.edu.cn

Abstract: *This paper studies a robot shuttle system (RSS), featured by automated guided vehicles (AGV) transporting storage totes with products in batches to order pickers. A robot shuttle system is a new type of automated material handling system, where products are stored in storage totes, so called totes-to-picker system. We develop a semi-open queueing network model (SOQN) to describe the RSS. The model can be used to effectively estimate system performance in terms of maximum order throughput capacity, order throughput time and resource utilization. Simulation experiments are conducted to validate the analytical model. We then conduct numerical experiments to investigate how the service batch size and the number of AGVs affect system performance. Through experimental results analysis, we provide guidelines on the optimization of these system design and decision related parameters.*

Keywords: *Robot shuttle system, Automated guided vehicles, Material handling system, Queueing network*

1 Introduction

In last decade, the percentage share of e-commerce sales has shown steady growth. Online retailers are facing the challenge of improving delivery speed and reducing overall system costs. Warehouse automation is considered as an effective solution to meet these requirements and has become a popular research topic in material handling ((Baker and Halim, 2007). Since order picking is one of the most critical tasks in warehousing, many automated order picking systems are developed in recent years, such as the robotic mobile fulfillment system (RMFS), and the automated vehicle storage and retrieval system (AVS/RS). The RMFS obtains high flexibility and scalability since it is deployed on the ground and employs movable shelves. However, the RMFS has poor space utilization due to the height limitation of movable shelves, which leads to higher storage costs. Moreover, robots in RMFS do lots of useless delivery. Robots in RMFS transport a pod at a time, while generally only a few products on the pod are required by an order. The robot shuttle system (Figure 1) is a new type of automated order picking system using innovative devices, which can address the above-mentioned issues. In an RSS, dense racks and storage totes (Figure1-b) are used to store products. Besides, it employs picking AGVs equipped with lifts to transport storage totes. As shown in Figure 1-c, there are several storage units and a lift on the AGV. This allows vehicles to transport only what is required according to customer orders. The RSS has been brought to market by companies such as Hai Robotics, Geek+, and Guozi Robotics. It has seen successful implementation in the famous office supplies seller, the Staples.

Compared to RMFS (e.g. Kiva system), the RSS has three main advantages: First, it has higher storage capacity and space utilization due to the use of dense racks and storage totes; Second, the picking vehicles in an RSS transport storage totes in batches to order pickers for order picking, which reduces the times of vehicles' round trips between storage area and workstations. Besides, the batch size can be varied according to real order picking demands. Therefore, the batch service of picking vehicles improve their handling efficiency and operational flexibility.

Third, a tote in RSS only contains one Stock keeping unit (SKU), and picking vehicles only need to bring which are required by customer orders to picking stations. This reduces average picking time of order pickers since they don't need to find required products from a whole pod and vehicles do less useless transport, and thus improving overall order picking efficiency. The disadvantages of the RSS are that the price of the advanced picking AGVs are much higher than a kiva robot, and the application of dense storage racks reduces system scalability.

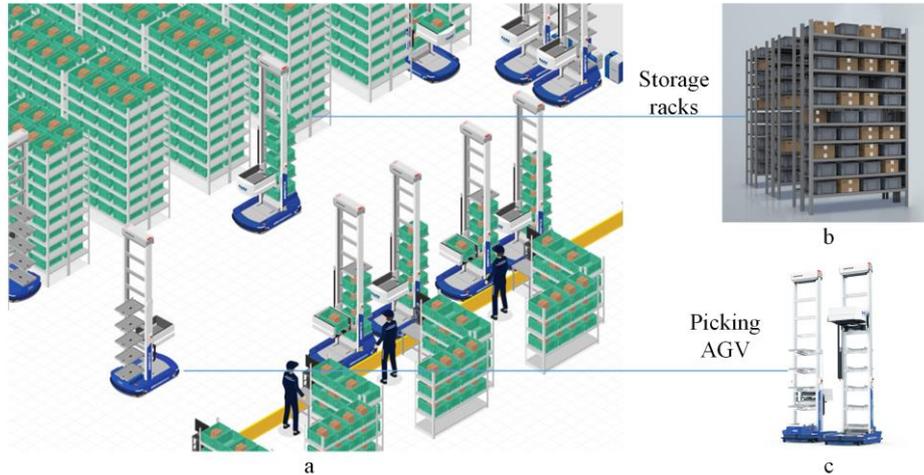


Figure 1: Illustration of a Robot Shuttle System

This paper focuses on the system design and decision optimization of the RSS. For this purpose, an analytical model based on queueing theory is developed for system performance estimation. The analytical model allows us to evaluate system performance under different system configurations and operation decisions with little computation time, which supports warehouse designers and managers to identify optimized system design and make appropriate operation decisions. Furthermore, through numerical experiments, this paper studies the following design and operation decision related research questions:

- (1) How does the batch size of picking AGVs affect system performance?
- (2) How does the number of AGVs affect system performance?

The remainder of the paper is organized as follows: section 2 reviews the literature. Section 3 introduces the robot shuttle system and the system work flow in detail. Section 4 presents the analytical model. Section 5 provides simulation results and section 6 provides numerical experiments and analysis. In section 7, we draw conclusions and provide future research directions.

2 Literature review

The RSS is a new type of order picking system that is derived from RMFS with certain technologies innovation. Due to limited studies on this new system, here we mainly review literatures on RMFSs, which can be divided into two categories, i.e. system design and operational decisions.

Design and analysis of robotic order picking systems is an attractive topic in light of considerable increase in online retails. Enright et al. (2011) described some allocation problems, such as pod storage allocation and order allocation in the RMFS to encourage future researchers to investigate it. Öztürkoğlu et al. (2019) proposed a new design idea, i.e. changing the angle of cross aisle, to find better layouts for RMFS. Simulation is a valuable tool to help evaluating system performance during system design. Lienert et al. (2018) presented a simulation model

for RMFS performance analysis, and the experiment results show a linear correlation between the number of vehicles and the throughput for small number of vehicles. Merschformann et al. (2019) analyzed the pod storage assignment and order assignment problems using discrete event simulation. Hanson et al. (2018) provided insights into the performance of RMFS and how it relates to the system design as well as the implementation context. Researchers also contribute to develop analytical models (e.g. queueing models) to analyze order picking systems. Yuan et al. (2017) built open network models for RMFS which can be used in the design of robotic warehouses. Guan et al. (2018) formulated an integer programming model to study the pod layout problem in RMFS, and a three-stage algorithm on the basis of the Spectral Clustering algorithm is proposed to solve the problem.

As for the operational decisions, Xiang et al. (2018) aimed at minimizing the number of visits of pods, by optimizing system storage assignment and order batching rules, thus reducing the useless traveling of robots in RMFS. Zoning strategies are also popular for optimizing RMFS storage assignment. Lamballais et al. (2020) optimized three decision variables of the RMFS by introducing a cross-class matching multi-class Semi-Open Queueing Network (SOQN). Zou et al. (2017) built a SOQN and a two-phase approximate approach for RMFS performance estimation. They proposed a near optimal order assignment rule based on handling speeds of workstations. Nils et al. (2017) focused on the order processing in a picking station and investigate the batching and sequencing strategies of picking orders. The results show that the optimized order picking allows to more than halve the fleet of robots. The robot allocation is another key factor which may influence performance of RMFS. Zhang et al. (2019) modeled this problem as a resource-constrained project scheduling problem, considering driving behavior of robots. Then a building-blocks-based genetic algorithm is proposed to solve this problem which is validated to be better than several classic and competitive crossover operators.

Although there are plenty of research focusing on RMFS, it is impossible to employ all the research conclusions on RSS, since that the RSS deploys innovative picking robots which works in totally different way. To the best of our knowledge, this paper is the first attempt to study RSS. Taking inspiration from the existing research on RMFS, a queueing network model which includes accurate driving behavior of vehicles is built to evaluate performance of RSS. We combine the batch service of vehicles into the model and investigate how the batch size influence system performance by measuring order throughput time and maximum order throughput capacity. The study of this paper provides guidelines for warehouse developers on optimizing both system design and operational decisions in practical application.

3 Robot shuttle system

This section provides comprehensive description of the robot shuttle system firstly. Then in 3.2.2 the picking process of the RSS is described in detail.

3.1 System description

A top view of the RSS layout is shown in Figure 2. The storage area consists of single-deep, double-sided storage racks, and these storage racks are divided into rectangular blocks by aisles and cross aisles. Each block is formulated by “ $2 \times$ ” storage racks, where 2 is the number of rows of storage racks. The picking area are situated at both ends of the storage area, where workstations and order walls are deployed. There is one picker at a workstation performing order picking, and the picked items are put into corresponding totes on the order wall according to customer orders.

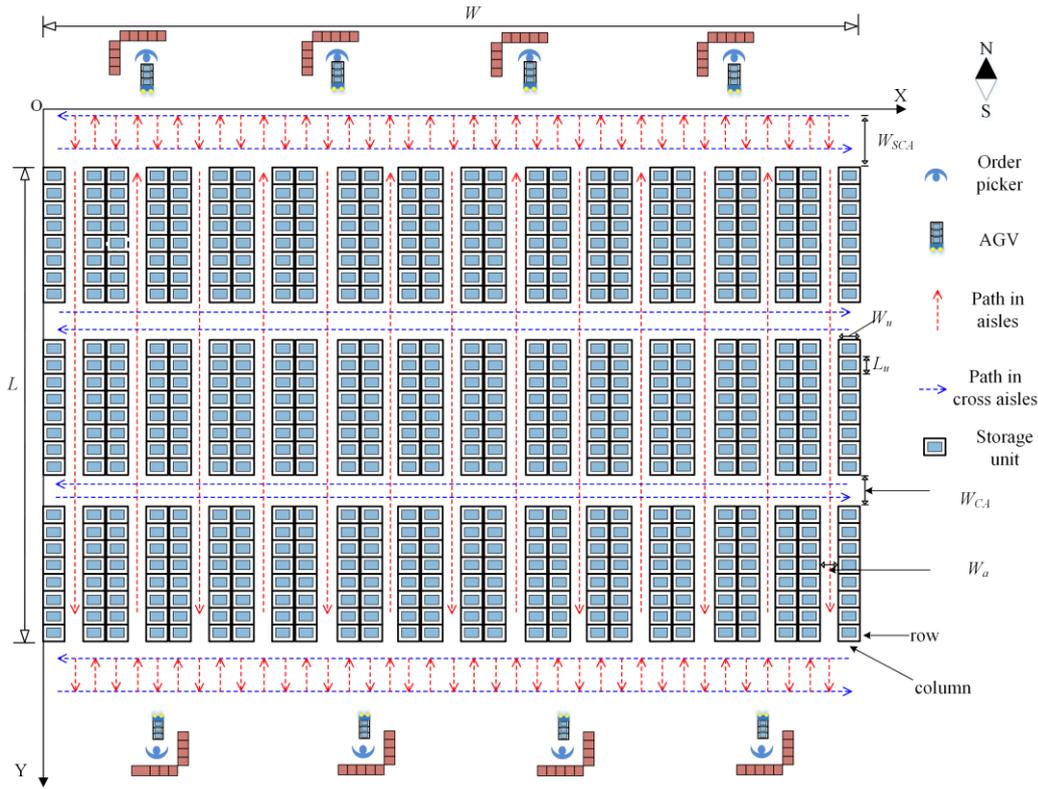


Figure 2: Layout of a Robot Shuttle System

Storage totes are stored on storage racks and each tote is stored at a dedicated storage unit. Each storage unit on racks can be indexed by its row number , column number and layer number , and contains one storage tote with one stock keeping unit. A coordinate system is formulated, where X-axis and Y-axis are arranged along cross aisles and aisles respectively. Then the column number increases with X-axis and the row number increases with Y-axis. The storage unit number can be calculated by

$$N_u = N_R \cdot N_L \cdot (c_u - 1) + N_R \cdot (r_u - 1) + l_u \quad (1)$$

Where N_R and N_L are the total number of rows and layers respectively.

The path planning in the proposed RSS is illustrated in Figure 2. All paths in both aisles and cross aisles are uni-directional to avoid congestion and deadlock. Considering the fact that picking AGVs in an RSS need to frequently travel through different aisles to perform batch service, we design two opposite paths (blue dotted lines) in each cross aisle to decrease travel distance and improve delivering efficiency, while each aisle only has a single path (red dotted lines) to improve space utilization. To simplify the calculation of vehicle moving time, we assume that two paths in a cross aisle are located at the middle of the aisle.

The main notations used in this paper are described in Table 1.

3.2 System workflow

The RSS fulfill order lines in batches, which means that picking AGVs transport a number of totes to a workstation at a time and then the picker picks required products from these totes. Therefore, the workflow is different from that in an RMFS. The main workflow is illustrated in Figure 3 and explained as follows:

Table 1: Notations in the paper

Notations	Description	Notations	Description
N_A, N_{CA}	number of aisles and cross aisles	N_C, N_R, N_L	number of columns, rows and layers
W_u, L_u	width and length of a storage unit(m)	W_A, L_{CA}	width of aisles and cross aisles (m)
W	total width of the storage area (m) $W = N_c \cdot W_u + N_A \cdot W_A$	L	total length of the storage area (m) $L = N_R \cdot L_u + N_{CA} \cdot W_{CA}$
N_{total}	number of storage units $N_{total} = N_R \cdot N_C \cdot N_L$	W_{SCA}	width of cross aisles between workstations and picking area (m)
N_W	Number of workstations	N_{WR}	Number of AGVs that served by one workstation
N	vehicle service batch size	N_l	Order lines in customer orders
τ_p	Picking time of order pickers (s)	τ_{lu}	Vehicles' loading/unloading time (s)
V	Average speed of vehicles (m/s)	τ_t	Turning time of vehicles
R_b	Number of rows in a storage rack block		

- ① The customer orders are split into order lines and wait in the external queues. When an AGV is released from last transport task, a number of order lines are assigned to the idle vehicle following a first-come-first-serve policy.
- ② Move 1: The AGV starts from its current position, and moves to each target tote in sequence, loading all the target totes. The process is described in Figure 3-a.
- ③ Move 2: When all the target totes in the order lines batch are loaded, the vehicle brings them to the claimed workstation, and waits in the queue if the picker is busy. The process is described in Figure 3-b.
- ④ Move 3: The picker picks up required products from totes on the vehicles, following a first-come first-serve policy. The picking time is stochastic, and we assume that it follows a general distribution, i.e. $\tau_p \sim U[a, b]$. After all the totes on the vehicle are handled, the vehicle transports the totes back to storage area, and the totes will be stored at their original storage units. The vehicle firstly moves from workstation to the first target storage unit, see Figure 3-c.
- ⑤ Move 4: Then the vehicle moves to each target tote in sequence, unloading all the target totes. The process is illustrated in Figure 3-d.
- ⑥ The main assumptions and operational rules are listed as follows: (1) In the RSS, we assume that retrieval task occurs at a random storage unit. (2) This study only considers order picking process while replenishment process is not considered, since the order picking strictly relates to service level. Thus we assume that there are always sufficient products to satisfy incoming order line, and no product shortage happens. (3) The picking AGVs are served by their dedicated workstations. (4) Congestion and deadlock may never happen due to the uni-directional paths applied. (5) The customer orders with different sizes arrive following a Poisson distribution with parameter λ . (6) The number of order

lines in an order is stochastic and follows a uniform distribution, $N_l \sim U[n_{min}, n_{max}]$. (7) Vehicles' loading/unloading time τ_{lu} is constant. (8) Vehicles transport a fixed number of totes, i.e. the batch size N is a constant.

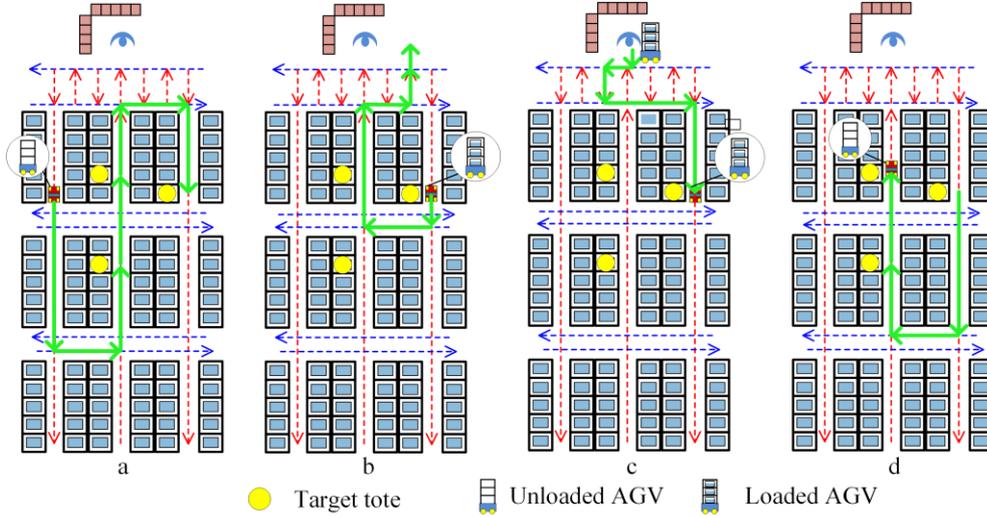


Figure 3: Illustrating the Main Workflow in the Robot Shuttle System

4 Analytical model for RSS performance estimation

Section 4.1 provides methods to calculate service time of vehicles. Section 4.2 presents a SOQN to estimate system performance. In section 4.3, a solution to solve the SOQN is described in detail.

4.1 Service time of vehicles

During the whole order picking process, vehicles' movements consist of four part, i.e. move 1 to move 4. Move 1 and move 3 are similar, which describe the travel between a workstation and an arbitrary storage unit. Move 2 and move 4 are also similar, which describe the travel among different storage units and the loading/unloading of vehicles. Since the loading/unloading time is constant, we only need to calculate the travel times of vehicles.

Firstly, we should know the location of each workstation and storage unit. As for the workstations, this study only considers the ones located on the northside of the storage area. We assume that the workstations are located on X-axis, and the index $i \in [1, N_w]$ increases along the X-axis. Then the coordinates of the i^{th} workstation can be denoted as:

$$(x_w, y_w) = \left(\left(i - \frac{1}{2} \right) \cdot \frac{W}{N_L}, 0 \right) \quad (2)$$

As for the storage units, since vehicles only move on the ground and perform loading/unloading on paths of aisles, the coordinates of a storage unit can be calculated by (3) according to the aisle number a_u and the row number r_u .

$$(x_u, y_u) = \left((2a_u - 1) \cdot W_u + \left(a_u - \frac{1}{2} \right) \cdot W_A, W_{CA} + \left(r_u - \frac{1}{2} \right) \cdot L_u \right) \quad (3)$$

In particular, we use y_b to denote the index of storage rack block that a storage unit belongs to along the Y-axis. y_b can be calculated according to r_u by (4):

$$y_b = \frac{r_u}{R_b} + 1 \quad (4)$$

Then we can calculate the travel time of vehicles, including the travel time from a workstation to a storage unit, and the travel time from a storage unit to another storage unit.

4.1.1 Travel time from a workstation to a storage unit

Assume that the vehicle starts from a storage unit (a_u, r_u, y_b) to the i^{th} workstation, the travel time is denoted by $\tau_{sw,i}$.

(1) If the path direction in the aisle is south, then the travel is composed of four movements.

① The vehicle moves to the nearest cross aisle along the path, the distance can be calculated by (5).

$$d_1 = \frac{W_{ca}}{2} + (y_b - r_u + \frac{1}{2}) \cdot L_u \quad (5)$$

② The vehicle moves towards the target workstation to the adjacent aisle, the distance is:

$$d_2 = W_{ca} + 2W_u \quad (6)$$

③ The vehicle moves to the cross aisle between the storage area and the picking area. The travel distance is:

$$d_3 = (y_b \cdot r_b - \frac{1}{2}) \cdot L_u + (y_b - 1) \cdot W_{ca} + \frac{W_{sca}}{2} \quad (7)$$

④ The vehicle moves to the workstation along the cross aisle, and the travel distance is:

$$d_4 = \left| x_{w,i} - x_u \right| - (W_a + 2W_u) \quad (8)$$

Then the total travel time can be calculated by equation (9).

$$\tau_{sw,i} = \frac{d_1 + d_2 + d_3 + d_4}{V} \quad (9)$$

(2) If the path direction in the aisle is north, then the travel is composed of two movements.

① The vehicle moves to the cross aisle between the storage area and the picking area. The travel distance is:

$$d_1 = (r_u - \frac{1}{2}) \cdot L_u + (y_b - 1) \cdot W_{ca} + \frac{W_{sca}}{2} \quad (10)$$

② The vehicle moves to the workstation along the cross aisle, and the travel distance is:

$$d_2 = (r_u - \frac{1}{2}) \cdot L_u + (y_b - 1) \cdot W_{ca} + \frac{W_{sca}}{2} \quad (11)$$

Then the total travel time can be calculated by equation (12).

$$\tau_{sw,i} = \frac{d_1 + d_2}{V} \quad (12)$$

4.1.2 Travel time between two storage units

Assume that the vehicle starts from a storage unit $(a_{u,1}, r_{u,1}, y_{b,1})$ to another storage unit $(a_{u,2}, r_{u,2}, y_{b,2})$, the travel time is denoted by τ_{ss} .

(1) When $a_{u,1} = a_{u,2}$, then the travel distance from $(a_{u,1}, r_{u,1}, y_{b,1})$ to $(a_{u,2}, r_{u,2}, y_{b,2})$ is:

$$d_1 = |r_{u,1} - r_{u,2}| \cdot L_u + |y_{b,1} - y_{b,2}| \cdot W_{ca} \quad (13)$$

Then the total travel time can be calculated by equation (14).

$$\tau_{ss} = \frac{d_1}{V} \quad (14)$$

(2) When $a_{u,1} \neq a_{u,2}$, if the path direction in $a_{u,1}$ is south and the path direction in $a_{u,2}$ is north, then the travel is composed of three movements.

① If $y_{b,1} < y_{b,2}$, then the vehicle moves to the nearest cross aisle on the south of block $y_{b,2}$; If $y_{b,1} \geq y_{b,2}$, then the vehicle moves to the nearest cross aisle along the path in $a_{u,1}$.

$$d_1 = \begin{cases} \left((y_{b,1} + |z_1 - z_2|) \cdot R_b - r_{u,1} + \frac{1}{2} \right) \cdot L_u & y_{b,1} < y_{b,2} \\ + \left(\frac{1}{2} + |y_{b,1} - y_{b,2}| \right) \cdot W_{ca}, & \\ \left(y_{b,1} \cdot R_b - r_{u,1} + \frac{1}{2} \right) \cdot L_u + \frac{W_{ca}}{2}, & y_{b,1} \geq y_{b,2} \end{cases} \quad (15)$$

② The vehicle moves to $a_{u,2}$, and the travel distance is:

$$d_2 = |x_{u,1} - x_{u,2}| \quad (16)$$

③ The vehicle moves to the location of the target storage unit along $a_{u,2}$, and the travel distance is:

$$d_3 = \begin{cases} \left(y_{b,1} \cdot R_b - r_{u,2} + \frac{1}{2} \right) \cdot L_u + \frac{W_{ca}}{2} & y_{b,1} < y_{b,2} \\ \left(y_{b,1} \cdot R_b - r_{u,2} + \frac{1}{2} + |y_{b,1} - y_{b,2}| \cdot R_b \right) \cdot L_u & \\ + \left(\frac{1}{2} + |y_{b,1} - y_{b,2}| \right) \cdot W_{ca}, & y_{b,1} \geq y_{b,2} \end{cases} \quad (17)$$

Then the total travel time can be calculated by equation (18).

$$\tau_{ss} = \frac{d_1 + d_2 + d_3}{V} \quad (18)$$

(3) When $a_{u,1} \neq a_{u,2}$, if the direction of paths in $a_{u,1}$ and $a_{u,2}$ are both south and $y_{b,1} < y_{b,2}$, then the travel is composed of three movements.

- ① The vehicle moves to the nearest cross aisle along the path in $a_{u,1}$, and the travel distance is:

$$d_1 = \left(y_{b,1} \cdot R_b - r_{u,1} + \frac{1}{2} \right) \cdot L_u \quad (19)$$

- ② The vehicle moves to $a_{u,2}$, and the travel distance is:

$$d_2 = |x_{u,1} - x_{u,2}| \quad (20)$$

- ③ The vehicle moves to the location of the target storage unit along $a_{u,2}$, and the travel distance is:

$$d_3 = \left(r_{u,2} - \frac{1}{2} + (|y_{b,1} - y_{b,2}| - y_{b,2}) \cdot R_b \right) \cdot L_u + \frac{W_{ca}}{2} \quad (21)$$

Then the total travel time can be calculated by equation (22).

$$\tau_{ss} = \frac{d_1 + d_2 + d_3}{V} \quad (22)$$

- (4) When $a_{u,1} \neq a_{u,2}$, if the direction of paths in $a_{u,1}$ and $a_{u,2}$ are both south and $y_{b,1} \geq y_{b,2}$, then the travel is composed of five movements.

- ① The vehicle moves to the nearest cross aisle along the path in $a_{u,1}$, and the travel distance is:

$$d_1 = \left(y_{b,1} \cdot R_b - r_{u,1} + \frac{1}{2} \right) \cdot L_u \quad (23)$$

- ② The vehicle moves towards $a_{u,2}$ to the adjacent aisle, and the distance is:

$$d_2 = W_a + 2W_u \quad (24)$$

- ③ The vehicle moves to the nearest cross aisle on the north of block $y_{b,2}$, and the travel distance is:

$$d_3 = (|y_{b,1} - y_{b,2}| + 1) \cdot (W_{ca} + R_b \cdot L_u) \quad (25)$$

- ④ The vehicle moves to $a_{u,2}$, and the travel distance is:

$$d_4 = (|y_{b,1} - y_{b,2}| - 1) \cdot (W_a + 2W_u) \quad (26)$$

- ⑤ The vehicle moves to the location of the target storage unit along $a_{u,2}$, and the travel distance is:

$$d_5 = \left(r_{u,2} - (y_{b,2} - 1) \cdot R_b - \frac{1}{2} R_b \right) \cdot L_u + \frac{W_{ca}}{2} \quad (27)$$

Then the total travel time can be calculated by equation (28).

$$\tau_{ss} = \frac{d_1 + d_2 + d_3 + d_4 + d_5}{V} \quad (28)$$

- (5) When $a_{u,1} \neq a_{u,2}$, if the path direction in $a_{u,1}$ is south and the path direction in $a_{u,2}$ is north,

then the travel is composed of three movements.

- ① If $y_{b,1} > y_{b,2}$, then the vehicle moves to the nearest cross aisle on the north of block $y_{b,2}$; If $y_{b,1} \leq y_{b,2}$, then the vehicle moves to the nearest cross aisle along the path in $a_{u,1}$. The travel distance can be calculated by (29).

$$d_1 = \begin{cases} \left(r_{u,1} - (y_{b,1} - 1 + |y_{b,1} - y_{b,2}|) \cdot R_b - \frac{1}{2} \right) \cdot L_u & y_{b,1} < y_{b,2} \\ + \left(\frac{1}{2} + |y_{b,1} - y_{b,2}| \right) \cdot W_{ca}, & \\ \left(r_{u,1} - (y_{b,1} - 1) \cdot R_b - \frac{1}{2} \right) \cdot L_u + \frac{W_{ca}}{2}, & y_{b,1} \geq y_{b,2} \end{cases} \quad (29)$$

- ② The vehicle moves to $a_{u,2}$, and the travel distance is:

$$d_2 = |x_{u,1} - x_{u,2}| \quad (30)$$

- ③ The vehicle moves to the location of the target storage unit along $a_{u,2}$, and the travel distance is:

$$d_3 = \begin{cases} \left(r_{u,2} - (y_{b,2} - 1) \cdot R_b - \frac{1}{2} \right) \cdot L_u + \frac{W_{ca}}{2}, & y_{b,1} > y_{b,2} \\ \left(r_{u,2} + \left(|y_{b,1} - y_{b,2}| - y_{b,1} + 1 - \frac{1}{2} \right) \cdot R_b - \frac{1}{2} \right) \cdot L_u & \\ + \left(\frac{1}{2} + |y_{b,1} - y_{b,2}| \right) \cdot W_{ca} & y_{b,1} \geq y_{b,2} \end{cases} \quad (31)$$

Then the total travel time can be calculated by equation (32).

$$\tau_{ss} = \frac{d_1 + d_2 + d_3}{V} \quad (32)$$

- (6) When $a_{u,1} \neq a_{u,2}$, if the direction of paths in $a_{u,1}$ and $a_{u,2}$ are both south and $y_{b,1} > y_{b,2}$, then the travel is composed of three movements.

- ① The vehicle moves to the nearest cross aisle along the path in $a_{u,1}$, and the travel distance is:

$$d_1 = \left(r_{u,1} - (y_{b,1} - 1) \cdot R_b - \frac{1}{2} \right) \cdot L_u + \frac{W_{ca}}{2} \quad (33)$$

- ② The vehicle moves to $a_{u,2}$, and the travel distance is:

$$d_2 = |x_{u,1} - x_{u,2}| \quad (34)$$

- ③ The vehicle moves to the location of the target storage unit along $a_{u,2}$, and the travel distance is:

$$d_3 = \left((|y_{b,1} - y_{b,2}| + y_{b,1} - 1) \cdot R_b - r_{u,2} + \frac{1}{2} \right) \cdot L_u + \left(|y_{b,1} - y_{b,2}| - \frac{1}{2} \right) \cdot W_{ca} \quad (35)$$

Then the total travel time can be calculated by equation (36).

$$\tau_{ss} = \frac{d_1 + d_2 + d_3}{V} \quad (36)$$

(7) When $a_{u,1} \neq a_{u,2}$, if the direction of paths in $a_{u,1}$ and $a_{u,2}$ are both south and $y_{b,1} \leq y_{b,2}$, then the travel is composed of five movements.

① The vehicle moves to the nearest cross aisle along the path in $a_{u,1}$, and the travel distance is:

$$d_1 = \left(r_{u,1} - (y_{b,1} - 1) \cdot R_b - \frac{1}{2} \right) \cdot L_u + \frac{W_{ca}}{2} \quad (37)$$

② The vehicle moves towards $a_{u,2}$ to the adjacent aisle, and the distance is:

$$d_2 = W_a + 2W_u \quad (38)$$

③ The vehicle moves to the nearest cross aisle on the south of block $y_{b,2}$, and the travel distance is:

$$d_3 = (|y_{b,1} - y_{b,2}| + 1) \cdot (W_{ca} + R_b \cdot L_u) \quad (39)$$

④ The vehicle moves to $a_{u,2}$, and the travel distance is:

$$d_4 = (|y_{b,1} - y_{b,2}| - 1) \cdot (W_a + 2W_u) \quad (40)$$

⑤ The vehicle moves to the location of the target storage unit along $a_{u,2}$, and the travel distance is:

$$d_5 = \left(y_{b,2} \cdot R_b - r_{u,2} + \frac{1}{2} \right) \cdot L_u + \frac{W_{ca}}{2} \quad (41)$$

Then the total travel time can be calculated by equation (42).

$$\tau_{ss} = \frac{d_1 + d_2 + d_3 + d_4 + d_5}{V} \quad (42)$$

4.2 SOQN for the robot shuttle system

The main objective of the paper is to formulate an analytical model for the RSS to estimate system performance, which can help us optimize system design and operation decision related parameters. Thus we construct a semi-open queueing network for the RSS, see Figure 4, and the SOQN model takes the batch service of vehicles into consideration. In the network, the order lines are assumed as customers. Since that the workstations work independently, we analyze the performance of a single workstation in isolation, and the analysis can be extended to other workstations similarly.

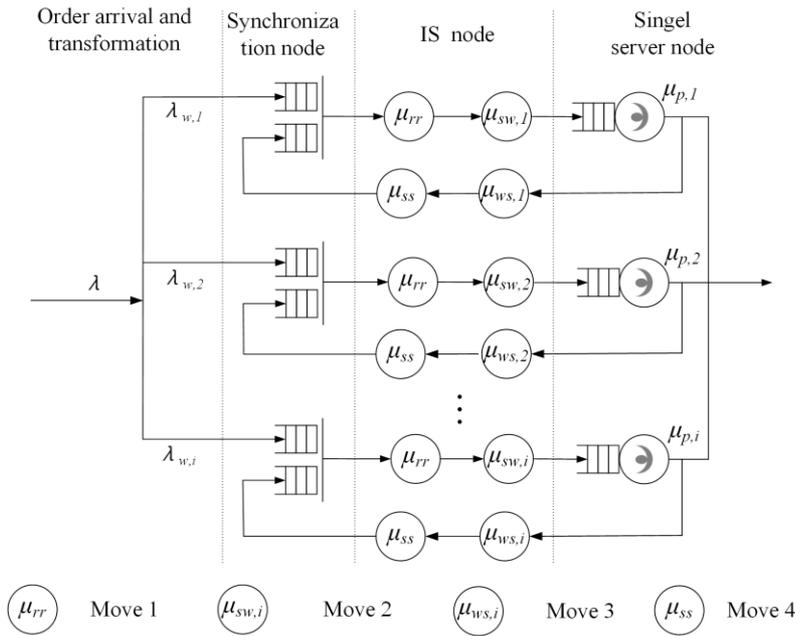


Figure 4: The Semi Open Queueing Network for the Robot Shuttle System

There are three kinds of server nodes in the proposed SOQN:

(1) Synchronization node.

When customer orders arrive at the system, they are split into individual order lines first, and then these order lines are paired with vehicles in batches. These procedures are implemted in the synchronization node, which consists of two queues, the queue of order lines Q_{ol} and the queue of vehicles Q_v .

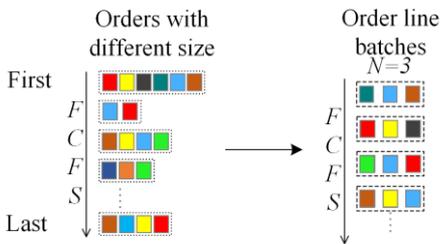


Figure 5: Illustrating the Transformation of Orders with Different Size to Order line batches

To analyze the performance of the synchronization node, we need to analyze the order arrival process first. As shown in Figure 5, the stream of arriving orders of different sizes are transformed to a stream of order line batches with batch size N . Assume that orders arrive at each workstation with identical probability, then the arrival rate of orders to the i^{th} workstation is $\lambda_{w,i} = \lambda / N_w$. Therefore, the arrival rate of order line batches to a workstation can be computed by (43).

$$\lambda_{b,i} = \frac{\lambda_{w,i} \cdot \overline{N}_l}{N} \tag{43}$$

Then the coefficient of variation $CV_{b,i}^2$ of interarrival time of order line batches with batch size N can be derived through the method by Bolch et al. (2006):

$$CV_{b,i}^2 = \frac{\overline{N}_l \cdot (CV_{w,i}^2 + CV_{N_l}^2)}{N} \quad (44)$$

Where $CV_{w,i}^2$ is the coefficient of variation of interarrival time of orders with different sizes, and $CV_{N_l}^2$ is the coefficient of variation of the orders' size.

(2) Infinite server (IS) node

When a batch of order lines are paired with an idle vehicle at the synchronization node, the vehicle can also be regarded as a customer. All the moves of the vehicle can be modelled as IS nodes since vehicles do not need to wait in any queue before moving, i.e. move 1, 2, 3, 4 are modelled by IS node u_{rr} , $u_{sw,i}$, $\mu_{ws,i}$, μ_{ss} respectively. In the model, the number of servers at an IS node are set as equal to the number of vehicles. The distribution of the service time of the IS nodes can be calculated based on the analysis in section 4.1, including the first and second moments of the service time.

(3) Single server node

At the workstation, vehicles wait in the queue and the order picker picks products from totes on the vehicles. Since there is only one picker at each workstation, the workstations are modelled as single server nodes. According to the distribution of the picking time, we can also calculate the first and second moments of the picking time.

The proposed SOQN model is analyzed using the solution procedure proposed by Buitenhek et al. in section 2.2. The maximum throughput TH_b , average throughput time OT_b and external queue L_o of the order line batches can be obtained. Then the maximum throughput TH_l , average throughput time OT_l and external queue $L_{q,l}$ of the order lines are calculated as follows:

$$TH_l = N \cdot TH_b \quad (45)$$

$$OT_l = OT_b \quad (46)$$

$$L_{q,l} = N \cdot L_o + \frac{N-1}{2} \quad (47)$$

Note that in (46), the first term represents the order lines in the L_o batches of the external queue, and the second term is the average number of order lines which are waiting to be combined into a batch.

5 Simulation validation

The discrete event simulation model is built with Arena (version 14.7), which complies with the real implementation of the RSS. The simulation model is assumed to conduct an infinite-horizon simulation, which can obtain a steady-state behavior analysis of the proposed robot shuttle system.

The simulation starts from an empty and idle state. In the simulation, the warm-up period of 10 hours is specified following the method by Welch (1983). The data collected during the warm-up period is disregarded to mitigate the presence of initialization bias. According to the rule of thumb in Banks et al. (2001), the simulation length is set as 100 hours, which is 10 times of the warm-up period. For each simulation, 30 replications are implemented.

The parameters for the simulation experiments are shown in Table 1. In the experiments, 9 scenarios are examined as a combination of three values of N and three values of N_{WR} , and three different order retrieval demand rates are set according to three different values of N_{WR} , to make the vehicle utilization no less than 60%.

N_{total}	N_C	N_R	N_L	N_{CA}	N_A	W_u	L_u	W_A	N_{WR}
16800	30	80	7	7	15	0.7 m	0.6 m	1 m	8, 10, 12
W_{CA}	W_{SCA}	τ_P	τ_{lu}	V_m	k_a	τ_t	N_l	R_b	N
2 m	2 m	$U[6s,10s]$	4s	2.1 m/s	3 m/s^3	1s	$U[1,5]$	10	3, 4, 5

The results are shown in Table 2. The system performance estimation results of simulation and analytical model are compared under 9 scenarios. The results show that, the deviation between analytical model and simulation are relatively low, which means that the proposed SOQN model can provide accurate estimation of system performance. Note that the OT_l estimated by the analytical model is always lower than that given by the simulation. This is mainly because that the time to combine order lines into batches is not considered in the analytical model. Besides, we can see that the system performance is affected by the service batch size N and the number of vehicles N_{WR} . Therefore, we investigate the influence of N and N_{WR} through numerical experiments in the next section.

Table 1: Estimation Results Comparison between Simulation (S) and Analytical Model (A)

N	N_{WR}	λ	OT_l		$L_{q,l}$		ρ_r		ρ_w	
			A	S	A	S	A(%)	S(%)	A(%)	S(%)
3	8	80	271.9	278.3	1.78	1.55	72.3	72.0	61.4	61.3
	10	100	291.6	297.1	1.99	1.72	77.7	77.9	76.4	76.3
	12	120	332.4	340.3	3.22	3.04	86.1	86.3	89.6	89.7
4	8	80	339.4	347.7	2.2	1.83	68.5	68.2	61.5	61.4
	10	100	368.2	375.4	2.38	2.06	74.5	75.1	76.6	76.6
	12	120	425.2	439.8	3.62	3.35	84.2	84.7	89.9	89.8
5	8	80	417.9	423.5	2.79	2.08	67.6	67.5	61.5	61.7
	10	100	454.7	470.1	3	2.39	73.8	74.1	76.6	76.5
	12	120	526.9	548.6	4.47	3.91	83.7	83.4	90	90.2

6 Numerical experiments

The number of vehicles N_{WR} should be determined during the system design phase, and the service batch size N is related to picking operation. We vary N and N_{WR} to study how they affect system performance, and five system performance measures are analyzed, namely: maximum throughput capacity TH_l , average throughput time OT_l , and external queue $L_{q,l}$ with respect to order lines, and the utilization of vehicles ρ_r and the picker ρ_w . The experimental scenario and basic parameters are the same as the simulation. In the experiment, two different

order retrieval demand rates are analyzed, namely: $\lambda_{w,i} = 40$ orders/hour and $\lambda_{w,i} = 80$ orders/hour. In each case, five different N values (i.e. $N = 1, 2, 3, 4, 5$) combined with 8 different N_{WR} values (i.e. $N_{WR} = 5, 6, 7, 8, 9, 10, 11, 12$) are considered.

According to Lamballais et al. (2017), the maximum order throughput capacity is independent of order retrieval demand level but depends on the system design and operational decisions. We analyze the TH_b of the robot shuttle system first, by removing the synchronization node from the proposed SOQN model, and then the TH_l can be derived from TH_b . The results are shown in

Figure 6.

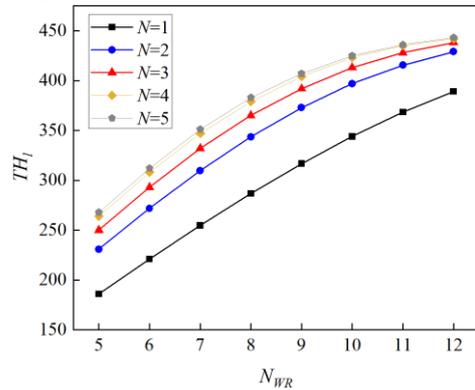


Figure 6: Illustration of TH_l Varying with N and N_{WR}

From the results we learn that, the TH_l increases with both N and N_{WR} . However, when the number of robots N_{WR} or the service batch size N exceeds a certain high level, the TH_l increases slightly with N_{WR} or N due to the limitation of the picking efficiency of the picker.

Therefore, warehouse designers should choose optimized values for N_{WR} and N to avoid over-productivity of the vehicles, which may help reduce overall system costs.

Then we analyze how the OT_l and the $L_{q,l}$ are affected by the N_{WR} and N . The results are shown in Figure 7 ($\lambda_{w,i} = 40$) and Figure 8 ($\lambda_{w,i} = 80$), where the utilization of vehicles ρ_r and the picker ρ_w are also presented. Three main conclusions we can draw from the results are:

- 1) When $\lambda_{w,i} = 40$, the utilization of vehicles ρ_r is relatively low. When $\lambda_{w,i} = 80$, the ρ_r is still maintained at a low level when a large number of vehicles are deployed, while the ρ_r is relatively high when there are fewer vehicles in the system.
- 2) From the results in the two figures, we learn that when ρ_r is maintained at a low level, both the OT_l and the $L_{q,l}$ are short and converge to certain values respectively with N_{WR} increasing. When ρ_r is high, increasing N_{WR} can decrease the OT_l and the $L_{q,l}$ effectively.

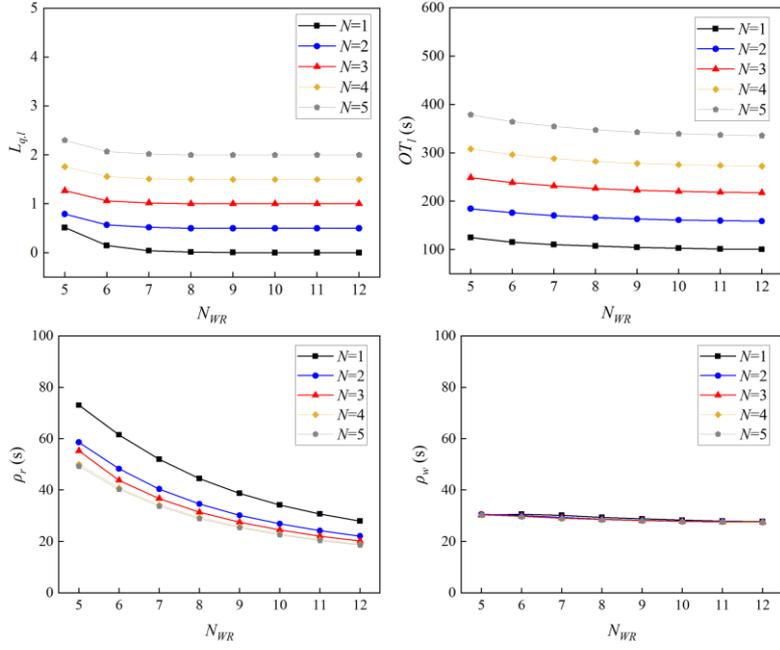


Figure 7: Illustration of System Performance Varying with N and N_{WR} ($\lambda_{w,i} = 40$)

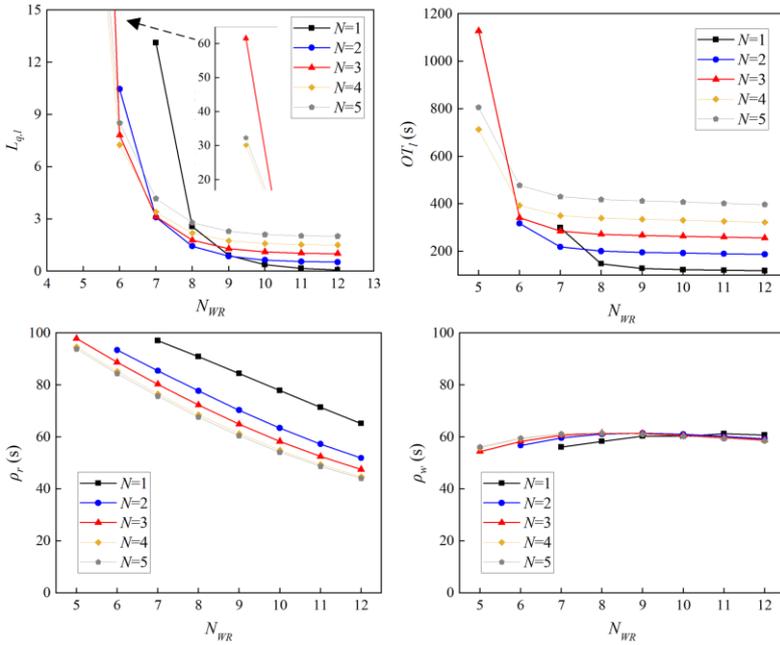


Figure 8: Illustration of System Performance Varying with N and N_{WR} ($\lambda_{w,i} = 40$)

- 3) When ρ_r is low, the main cause of external queue is that order lines need to be combined into a batch. According to the second term in formula (46), the $L_{q,i}$ will increase with N . We can see that the results with respect to $L_{q,i}$ validate the above analysis. The OT_i also increases with N since that the vehicles and the picker need to handle more order lines in a handling cycle.

When ρ_r is high, increasing N may improve the operational efficiency to a certain extent, which can decrease $L_{q,l}$, as well as the OT_l . If N is very low, (e.g. when $\lambda_{w,i} = 80$, $N_{WR} = 5$, and $N = 1$), the system cannot even reach a steady state.

Overall, increasing both N and N_{WR} reasonably according to order retrieval demand level can improve system performance. However, excessive increase of N_{WR} may cause over-productivity of vehicles. Similarly, over-increase of N may increase OT_l considerably and thus cause severe delay of order delivery.

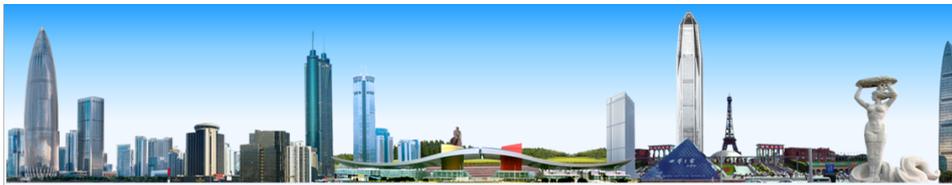
7 Conclusions

In this study, we focus on the performance analysis of robot shuttle system. First, a semi open queueing network model is developed to provide accurate performance estimation for the RSS. The effectiveness of the analytical model is confirmed by simulation experiments. The main implication of the analytical model is to help warehouse developers evaluate system performance under different system configurations efficiently. The study also provides guidelines for warehouse designers and managers on how to identify an appropriate service batch size and a proper number of AGVs within a workstation, which can avoid over-productivity of vehicles and lower system costs. In the future research, the impact of rack layouts on system performance may be taken into consideration.

References

- Banks J., J. S. C. II, B. L. Nelson, and D. M. Nicol (2001): *Discrete-event system simulation /-3rd edn.* Prentice Hall, 2001.
- Bolch G., S. Greiner, H. De Meer, and K. S. Trivedi (2006): *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*, 2ed ed. New Jersey: John Wiley & Sons, 2006.
- Baker P. and Z. Halim (2007): "An exploration of warehouse automation implementations: cost, service and flexibility issues," *Supply Chain Management: An International Journal*, vol. 12, no. 2, pp. 129-138, 20 March 2007.
- Boysen N., D. Briskorn, and S. Emde (2017): "Parts-to-picker based order processing in a rack-moving mobile robots environment," *Eur J Oper Res*, vol. 262, no. 2, 2, pp. 550-562, 2017.
- Enright J. J. and P. R. Wurman (2011) "Optimization and coordinated autonomy in mobile fulfillment systems," in *Proceedings of the 9th AAAI Conference on Automated Action Planning for Autonomous Mobile Robots*, San Francisco, California, USA, 2011, 2908681: AAAI Press, pp. 33-38.
- Guan M. and Z. Li (2018): "Pod Layout Problem in Kiva Mobile Fulfillment System Using Synchronized Zoning," *Journal of Applied Mathematics and Physics*, vol. 06, no. 12, pp. 2553-2562, 2018.

- Hanson R., L. Medbo, and M. I. Johansson (2018): "Performance Characteristics of Robotic Mobile Fulfillment Systems in Order Picking Applications," *IFAC-PapersOnLine*, vol. 51, no. 11, 11, pp. 1493-1498, 2018.
- Lamballais T., D. Roy, and M. B. M. De Koster (2017): "Estimating performance in a Robotic Mobile Fulfillment System," *Eur J Oper Res*, vol. 256, no. 3, 3, pp. 976-990, Feb 2017.
- Lienert T., T. Staab, C. Ludwig, and J. Fottner (2018): "Simulation-based Performance Analysis in Robotic Mobile Fulfillment Systems," in *Proceedings of the 8th International Conference on Simulation and Modeling Methodologies, Technologies and Applications*, Porto, Portugal, 2018: SciTePress, 2018, pp. 383-390.
- Lamballais T., D. Roy, and R. B. M. De Koster (2019): "Inventory allocation in robotic mobile fulfillment systems," *IISE Transactions*, vol. 52, no. 1, pp. 1-17, 2020/01/02 2020.
- Merschformann M., T. Lamballais, M. B. M. D. Koster, and L. Suhl, "Decision Rules for Robotic Mobile Fulfillment Systems," *Operations Research Perspectives*, vol. 6, pp. 100-128, 2019.
- Öztürkoğlu Ö. and A. Mağara (2019): "A New Layout Problem for Order-Picking Warehouses," in *Proceedings of the 9th international conference on industrial engineering and operations management*, Bangkok, Thailand, 2019, pp. 1047-1058.
- Welch P. D. (1983): "The statistical analysis of simulation results," *The computer performance modeling handbook*, vol. 22, pp. 268-328, 1983.
- Xiang X., C. Liu, and L. Miao (2018): "Storage assignment and order batching problem in Kiva mobile fulfillment system," *Engineering Optimization*, vol. 50, no. 11, 11, pp. 1941-1962, 2018.
- Yuan Z. and Y. Y. Gong (2017): "Bot-In-Time Delivery for Robotic Mobile Fulfillment Systems," *IEEE Transactions on Engineering Management*, vol. 64, no. 1, 1, pp. 83-93, 2017.
- Zhang J., F. Yang, and X. Weng (2019): "A Building-Block-Based Genetic Algorithm for Solving the Robots Allocation Problem in a Robotic Mobile Fulfillment System," *Mathematical Problems in Engineering*, vol. 2019, pp. 1-15, 2019.



Data-driven analytics-based capacity management for hyperconnected third-party logistics providers

Jana Boerger^{1,2,4} and Benoit Montreuil^{1,2,3,4}

1. Physical Internet Center

2. Supply Chain and Logistics Institute

3. Coca-Cola Chair in Material Handling & Distribution

4. H. Milton Stewart School of Industrial & Systems Engineering

Georgia Institute of Technology, Atlanta, USA

Corresponding author: jana@gatech.edu

Abstract: *In this paper we provide justifications why and ways how to enable 3PLs to be poised for success in the Physical Internet (PI) while facing a highly competitive and uncertain world. We notably argue that 3PLs have to transform from relying on static, inflexible, and disconnected ways and technologies for managing their capacity, to leveraging dynamic, flexible, and interconnected ways and technologies. Indeed, in the PI context, 3PLs have to be keen to achieve hyperconnectivity and manage capacities in multi-tenant warehouses more efficiently by leveraging data and ultimately increasing revenues and profits. We specifically propose a three-layer decision-making framework that offers 3PL organizations one stepstone enabling this transformation: successfully translating available data into decision-making, increasing service capabilities and performance, revenues and profitability, as well as sustainability. In the framework, a descriptive layer allows visibility over past capacity and activity related to key resources (e.g. storage capacity), a predictive layer allows visibility in the future, and a prescriptive layer allows automatic and dynamic diagnosis and planning to fully exploit and develop capacity and to best serve clients and the overall market. The framework maps descriptive, predictive, and prescriptive analytics to outcome-oriented activities, and to their data-driven and/or model-based foundations. The framework currently focuses on capacity management for warehousing, distribution, and fulfillment facilities, and can be expanded to encompass all logistics offers, activities, and assets of a 3PL as part of a logistics web. The contribution is illustrated through the context of a major American 3PL.*

Keywords: *3PL, Capacity Management, Data-driven Decision-Making, Decision Automation, Decision Support Systems, Demand Forecasting, Hyperconnected Logistics, Physical Internet, Supply Chain Management, Warehouse Management*

1 Introduction

The traditional 3rd party logistics provider (3PL) has long-term contracts with its customers, negotiated when existing contract terms come to an end, and when new aspiring to sign new customers. This 3PL is also very asset intensive, reaping revenues from owning assets and offering them to their customers for a fixed and typically long period of time. This traditional 3PL is well adapted to the world of past decades. Indeed, in a world that is only slowly changing, this traditional 3PL can be successful through its double focus on long-term selling and planning from one side, and on steady operational excellence from the other side.

Today however, the world is ever more characterized by volatility, uncertainty, complexity and ambiguity (VUCA, Bennett, 2014). In the logistics environment, VUCA's volatility and uncertainty induce a highly competitive market with companies having products with short

product life cycles and many promotions (Packowski, 2014), which then translates into high fluctuations in demand for logistic services and capacity. As depicted in Figure 1, these fluctuations result in situations where warehouses face a risk of overflowing, or capacity becomes available and remains unused, calling for improved capacity management.

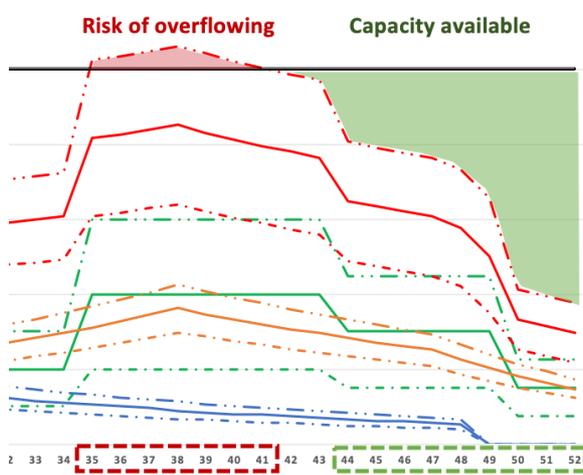


Figure 1: Impact of demand volatility on warehouse capacity

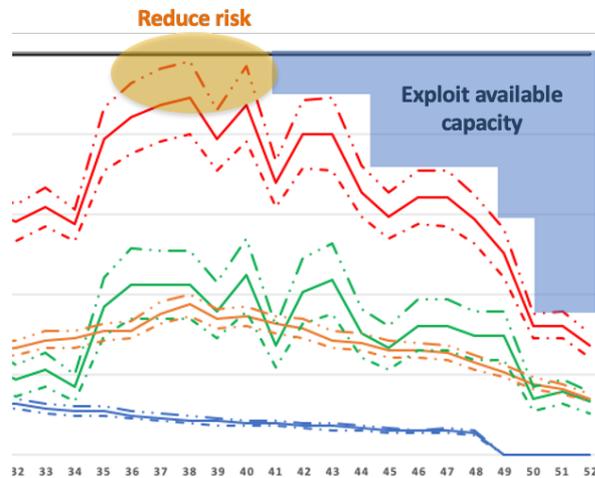


Figure 2: Successfully managed demand volatility

VUCA’s complexity is notably induced by the increasing product portfolio of clients and the increased pressure for reliable timeliness, resulting in a higher number of individual SKUs (stock keeping units) to be managed by warehouses in a fast-paced, often omnichannel context. This creates a high pressure environment for competitiveness, efficiency and sustainability for all logistics companies, and thus for logistics service provider. To become an advanced player in this context, the company needs to be able to dynamically manage its assets, countering the VUCA world with vision, understanding, clarity and agility so that it can reduce and manage risks, exploit available capacity, and develop capacity options (e.g. Figure 2). It can do so by adopting the hyperconnected paradigm through the Physical Internet (Montreuil, 2011; Montreuil et al., 2013; Ballot et al., 2014; Montreuil, 2017), with more dynamic and open interconnection with clients on one side, and with other logistics web players on the other side. Client interconnectivity enables higher information and communication capabilities, and dynamic elaboration of win-win service and capacity offers. Logistics player interconnectivity enables to enhance the services and capacity options that can be leveraged to smartly fulfill client needs.

Becoming an advanced hyperconnected logistics service provider in the VUCA Physical Internet world requires a full transformation along many threads. Our contribution lies in one of these required threads: the ability to manage 3PL capacity in a smart, dynamic, hyperconnected way.

As a key enabler for this transformation, we hereafter propose a three-layer decision-making framework that includes a descriptive layer, a predictive layer and a prescriptive layer. We argue that implementing and leveraging this analytics-based framework to build 3PL capability in logistics capacity management is a necessary step towards thriving in a VUCA Physical Internet world.

We first briefly review in section two the literature that has been published on 3PLs, their decision-making and analytic frameworks. We then outline in section three key differences between a traditional 3PL and a hyperconnected 3PL. In the fourth section, we propose our

data-driven capacity management decision-making framework to enable 3PLs to monitor, predict and plan their warehouse capacity. Note that the words “warehouse” and “facility” are used interchangeably throughout this paper, both naming a warehouse that the 3PL operates to serve its customers. Finally, in section five we provide conclusive remarks and avenues for further research.

2 Literature Review

Third-party logistics provider have an increasingly important role in today’s supply chains, becoming the core orchestrator of many companies’ supply chains. They therefore face a need to improve their efficiency and effectiveness (Zacharia, 2011). Despite these developments, to the best of our knowledge, there is no research focusing on capacity management for logistics service providers and their facilities.

Research concerning 3PLs is often (1) written from the point of view of other industry companies that are looking to use the services of 3PLs, (2) analyzing 3PL market development, or (3) analyzing the competitiveness of logistics providers. While these are observational studies, they fall short of proposing frameworks for 3PLs to work with. Hertz and Alfredsson (2003) analyze the development of companies that enter the field of 3PL business from being integrators, standard shipping firms or traditional brokers. Marchet et al. (2017) find that while 3PLs operate in a competitive market, only 25% of 3PLs are at the technical efficiency frontier and only 10% have innovative processes.

The notion of descriptive, predictive and prescriptive analytics has been discussed in the world of business analytics and in the context of supply chain analytics. Souza (2014) notably showcases that analytics is not new in supply chain management and that with the increasing amount of data available, opportunities for the application of analytics increase.

The research that is most related to our work is the framework developed by Hahn and Packowski (2015) for supply chain management. Their framework associates descriptive, predicative, and prescriptive analytic approach with types of use cases and methodological requirements from a business perspective, and with decision support systems concepts and formal types of IT systems from an information technology perspective.

Their four use case types are monitor-and-navigate, sense-and-respond, predict-and-act, and plan-and-optimize. The uses cases are associated by pairs to methodologies, respectively: monitoring and reporting, data modeling and mining, forecasting and simulation, strategic and operational planning. Descriptive analytics is mainly data-driven and relying on systems such as Enterprise Resource Planning (ERP) systems, expert systems, and business intelligence (BI) systems. Prescriptive analytics is mainly model driven, enabled by advanced planning systems (APS). Predictive analytics stands between them, borrowing from both model and data driven concepts, and relying on APS, BI, and expert systems.

Borrowing from Hahn and Packowski (2015), we adapt it to address the specific challenges of hyperconnected 3PLs and expand it to encompass the activities related to managing capacity in multi-tenant 3PL facilities.

3 Traditional 3PL vs Hyperconnected 3PL

In general, 3PLs may provide a variety of services to their customers, notably transportation, forwarding, warehousing, and value-adding services (VAS) such as relabeling/repackaging, assembly/installation, and blast freezing.

In this paper, we focus on 3PLs that own or lease, and operate, deep storage warehouses, distribution centers as well as fulfillment centers. The customers of these 3PLs are producers, distributors, and/or retailers (brick-and-mortar, e-commerce, and omnichannel). So, some of the customers are upstream in the supply chain while others are downstream. Traditionally, these 3PLs sign contracts with larger customers that tend to be long-term agreements that are renegotiated every three to five years. They often serve smaller customers on an as-needed basis, accommodating their small flow and storage of pallets and cases. Naturally, this leads to multi-tenant warehouse environments where multiple customers share one facility of the 3PL. The multi-tenant characteristic is the critical complexity factor justifying the emphasis on smart capacity management capabilities addressed in this paper.

Each customer has unique dynamic patterns relative to their inbound flow, storage needs, and outbound flow. The shock of these multiple customer-specific patterns can create significant disruptions, some positive, some negative, and some potentially both, yet all having to be addressed.

Relative to disruptions, consider for example a case where it becomes clear that a major customer tenant of a 3PL is to use significantly less storage space and throughput capacity than allowed in its contract, such as illustrated in Figure 2. Normally, its contract has it pay for the storage space, whether or not it uses that space, yet it is to be charged for operational inbound and outbound activities only if these actually occur. Negatively, this means that the 3PL is to have less revenues from that client, a fact attenuated somewhat as this client will require less resources to serve it, and thus induce less costs. Positively, this can be smartly turned into an opportunity if the 3PL recognizes fast enough the situation and is capable of offering to other customers the time-window-specific extra availability of space and throughput capacity, in a win-win mode for the customer tenant at the source of this opportunity.

Relative to risks needing to be managed, consider overflows as an example. Overflows happen when 3PLs, similarly as airlines with passengers, book more flow and storage than they are capable of dealing with concurrently, betting on the stochasticity to smooth requirements, or simply due to them not having planned their capacity commitments correctly. Overflows create havoc as excessive concurrent truck arrivals and excessive total goods inventory in a warehouse cause serious productivity disruptions with lack of available docks, too many trucks and trailers waiting in the yard and beyond, almost no available storage bays, overspill of stock in aisles, and huge congestion due to high flow intensity and disrupted aisles, potentially leading to an ultimate complete operational deadlock. Risks of overflowing need to be managed smartly. Indeed, 3PLs usually like tenants to use their allotted capacity at a high level inducing lucrative high inbound and outbound operations and revenues. Yet, when most tenants use near their maximal contractually allotted capacity, and some going overboard, there is significant risk of overpassing a threshold leading into overflow and deadlock. This risk and reward trade-off needs to be carefully managed.

A hyperconnected 3PL is to face the same challenges as traditional 3PLs, yet with higher intensity and dexterity. Let us consider first the intensity perspective. In the Physical Internet, the clients of logistics service providers aim to seamlessly deploy dynamically their products in a way enabling them to offer their customers fast, cheap, convenient, and reliable fulfillment services. They want to be able to shift products to locations best fitting the swiftly-changing market patterns, and to do so in an efficient and economical way. This leads them to request shorter and/or more flexible contracts, with less restrictive commitments blocking them from their aspirations toward best serving their customers. Also, the Physical Internet openly interconnects logistics networks, which induces each node of the overall logistics web to be

prepared to deal with more customers, as long as they respect and use the standardized protocols, interfaces and modular encapsulation. This means potentially more contracts of shorter duration with more distinct clients. Overall, this heightens the intensity of the capacity management challenges, requiring 3PLs to act according to a higher clock speed, and with more agility, adaptability, and resilience.

Let us now consider the dexterity perspective. In the Physical Internet, logistics service providers are to be interconnected much more and better on multiple layers, including physical, digital, operational, transactional, legal, and personal layers, with clients and other logistic service providers. This interconnection is not to be achieved solely through long-term contracts, alliances, and consortiums, but rather through accepting to act according to standardized protocols, leveraging standardized interfaces notably embedded in digital platforms and marketplaces, and using standardized modular containers across industries and across territories. The hyperconnected 3PLs are notably to be exchanging operational and transactional data on a much faster and intense pace with their clients and other logistics service providers used by their clients and/or offering capacity options leverageable for dealing with dynamic surges in capacity requirements. Exchanging plans and forecasts with clients, focused on their intersection space, is to be customary, enabling both to best anticipate and respond to forthcoming certain and uncertain changes. The same goes first, amongst the facilities and business units of a single logistics provider, and second, between hyperconnected logistics service providers. Each provider becomes a source of capacity options for the others, and everyone is part of the multi-service-provider supply web of multiple clients, having to interact to ensure smooth, seamless, and efficient overall performance. Overall this heightens the required dexterity of logistics service providers in meeting capacity management challenges, equipped with interconnected smart tools, and trained to think, plan and act in the Physical Internet so to achieve the necessary efficiency, agility, adaptability, and resilience.

The combination of heightened intensity and dexterity puts significant pressure on raising the capabilities of 3PLs for managing their capacity in a much more proactive way, fed by data from interconnected sources within their own organization and with interacting clients and other logistics providers, through direct links or platforms. As a contribution to this quest, the framework introduced in this paper guides the development of decision support technologies and processes for hyperconnected 3PL capacity management.

4 Data-driven capacity management decision-making framework

The decision-making framework, depicted in Figure 3, links three components: the type of analytics approach, namely descriptive, predictive, and prescriptive; the key groups of outcome-oriented activities; and the data-driven and/or model-based foundations. Each analytics approach is linked to a set of outcome-oriented activities, and each of these is calibrated in terms of its relative reliance on data-driven vs model-based foundations.

The decision-making framework has three layers of analytics approaches: the descriptive layer, the predictive layer and the prescriptive layer. This is line with the works of Hahn and Packowxski (2015), and as described in the landmark work of Davenport and Harris (2017) in the update to their work from 2007 that introduced business analytics. Some analytics professionals also argue that a fourth layer should be explicitly identified, that is diagnostic analytics, referring to the analysis of why something happened (e.g. Banerjee, 2013). In the framework, even though we recognize the importance of diagnostic analysis, we have not made it a fourth layer, but rather incorporated analytical diagnosis in each of the analytical layers. To predict future activity successfully (predictive layer), one needs to be aware of the underlying

factors that result in certain activities. At the descriptive layer, the reason for certain flow activities are a result of market movement. To understand the why of certain flows, market factors are therefore incorporated into the descriptive layer. For example, during the initial impact of the COVID-19 crisis in the USA, for the American 3PL storage usage changed. Some customers saw increased inventory, while others saw decreasing inventory not being able to keep up with the market. In the descriptive layer, it is not only sufficient to highlight the shifts of inventory, it is also important to help diagnose why these happening. Overall, in this example, the COVID crisis is a root cause, yet it must help to understand why some activities climbed while others went down, here notably linking with increasing demand on the market for essential products, and decreasing capacity in COVID affected supply chains. Seeking to raise alerts and to identify causes is at the core of the framework, at the three analytics layers.

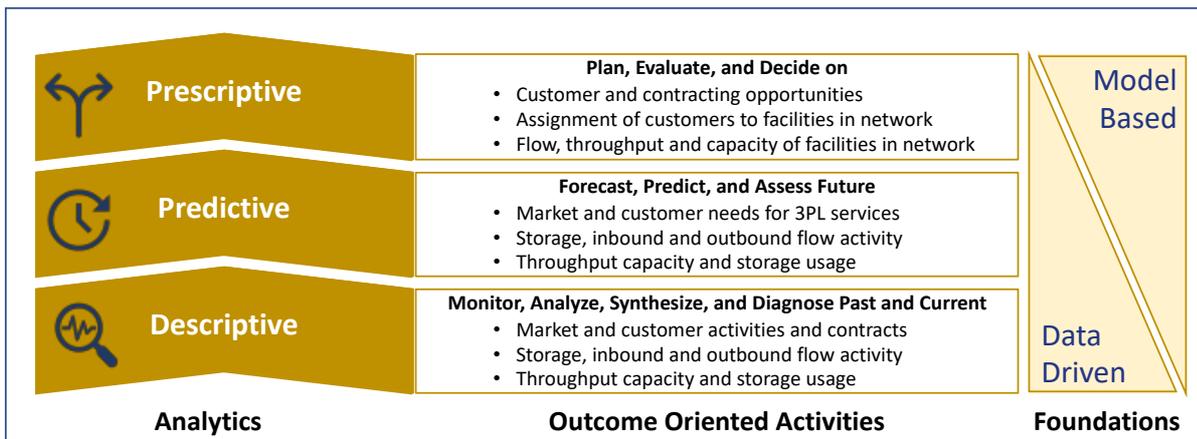


Figure 3: Data-driven model-based logistics provider capacity management framework

We hereafter describe the framework further by focusing on each of the three analytics layers and their outcome-oriented activities. Each layer concerns three aspects of the capacity management task: the market, the customer and its own network. We emphasize how the layers combine to allow an effective management of storage and throughput capacity in 3PLs’ facilities.

4.1 Descriptive Layer

The descriptive layer allows monitoring of the current activity of the overall market, of the facility network and at the level of an individual warehouse. It offers near real-time insights and visibility across its network to decision-makers. It is in line with the well-recognized importance of visibility as a core attribute of 3PLs, along being a neutral arbitrator and collaborator (Zacharia, 2011).

On a facility level, the descriptive layer must let a decision-maker see the current storage and throughput capacity available, current throughput demand, storage demand but also customer service level. More importantly, it should allow to highlight the usage of this capacity per combinations of warehouses and customers. Questions such as “How much capacity does Customer X currently use in our facilities?” should be easily answered overall and per facility.

Relative to monitoring capability, the descriptive layer should also allow a 3PL decision-maker to look at the historical development of the activities in specific facilities. In addition to usage, the descriptive layer should offer visibility over flows in the network: storage, inbound and outbound flows have to be monitored, and tracing should be kept over time.

Lastly, a fully implemented descriptive layer also allows visibility into the general 3PL market, customer activities and current contracts. Since this last aspect depends on outside information,

it is harder to implement in the early Physical Internet phases and thus, the initial focus of 3PLs is expected to be within the 3PL organization, and then gradually evolve to encompass this wide-angle out-of-the-box visibility.

Monitoring, analyzing, diagnosing, and synthesizing the current and historical activities provide the decision-maker facts, insights, intuition about the state of the activities and how they generally behave. Monitoring facts, states, and events is clearly data-driven. Analysis is fed by the monitored data, yet is often sustained by some high-level descriptive model to structure the approach. Diagnosis builds on monitoring and analysis, being strongly data-driven, yet often builds upon rule-based models and cause-and-effects models. Synthesis builds on monitoring, analysis and diagnosis, and is mostly still relying on human-centric skills combining reasoning, mental models, intuition, and discussions.

When attempting to move forward and decide on future actions, descriptive analytics sets the stage, yet it becomes critical to understand and project future capabilities and capacities, which is the focus of the predictive layer of the framework.

4.2 Predictive Layer

The predictive layer aims to offer reliable forecasts of forthcoming capacity and throughput demand and as a result capacity utilization, future service levels, and flows throughout the network. For prediction purposes, this layer builds upon hybrid timeseries forecasting methods (Zhang, 2003) based on traditional methods such as ARIMA and machine learning techniques (Ahmed, 2010) such as neural networks.

At the predictive layer, the power of the Physical Internet comes increasingly into play. Through hyperconnectivity with its customers, the 3PL may gain access to their current demand and/or supply logs and predictions. These predictions can include the customers' production plan and potentially privacy-protected point-of-sale (POS) data that it receives from its retailers, or the equivalent from e-commerce websites. This source offers richer data than the data generated internally by the 3PL, which represents solely its own history. In the predictive layer, the forecasts from the 3PL and the forecasts from the customer should then be ensembled into an overall forecast as depicted in Figure 4.

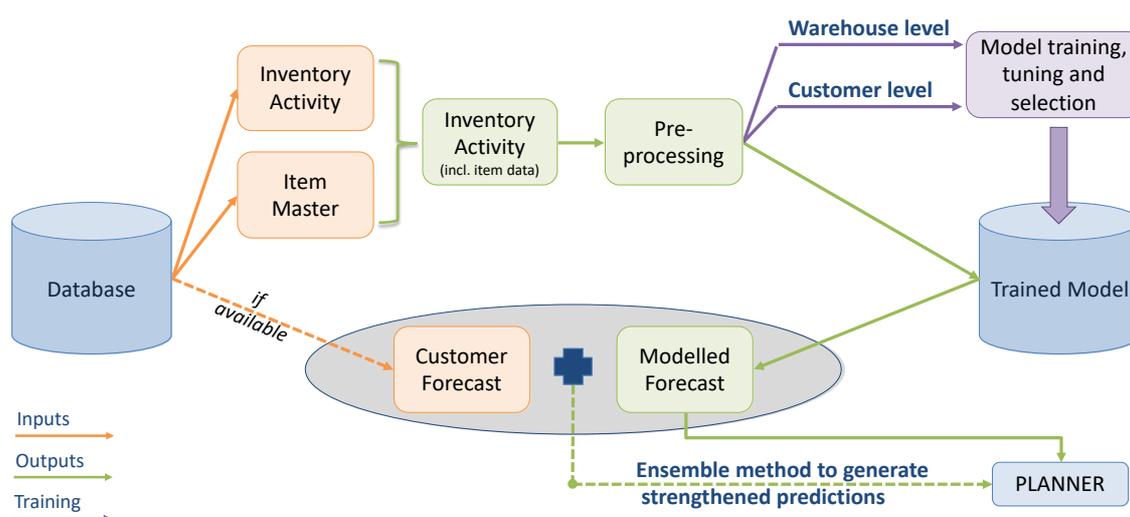


Figure 4: Possible information flow in predictive layer

Such ensembled forecasts, generated by combining several forecasts, have long been known to have the potential for better accuracy (e.g. Bates, 1969) and have become a core part of the

fields of ensemble learning and statistical learning (Hastie, 2009). In our initial experiments with the American 3PL, ensembled forecasts have shown as expected to have a higher accuracy, resulting in lower forecast errors.

Strengthened by the hyperconnectivity and the customers' forecasts, the predictive layer then should project expected future warehouse activity while explicitly recognizing the uncertainty in its prediction. It is important to note that it is usually impossible to reach 100% accuracy in predictions and it is thus important for the decision-maker to understand the accuracy reliability of a forecast. To support this understanding, uncertainties should be clearly exposed by the descriptive layer through prediction intervals, such as X% lower bound, most probable, and X% upper bound. As X climbs to higher levels, such as 99.9%, the prediction interval gets wider. It usually also gets higher as the future horizon covering the prediction is farther away (e.g. for tomorrow, next Monday, Thanksgiving) and usually gets relatively smaller with higher aggregation (for a specific day vs a week or a month). Explicitly acknowledging uncertainty and prediction accuracy is fundamental to assess correctly the forthcoming future and to enable well-informed decision making.

4.3 Prescriptive Layer

The prescriptive layer aims to offer decision facilitation capability. It builds upon the descriptive layer and the predictive layer, and enables decisions based upon the output of these layers. It is the final layer in the framework and offers the 3PL support in decisions concerning the market, the network and individual facilities. Activities such as accepting, rejecting and seeking customers and new contracting opportunities that fall into the area of business development are supported through the information available in the descriptive layer. It also helps to assign customers to facilities within the 3PL's network and can suggest potential assignment adaptations. In addition to these strategic and tactical activities concerning the customers, the prescriptive layer can also suggest adaptations of the flow, throughput and capacity of facilities in the network (Figure 3).

The prescriptive layer should help the 3PL to plan for future growth and contraction. Future growth of a customer might expand beyond the capacity available at a facility. To preclude related service failure, the 3PL can act proactively with the support of the prescriptive layer. It could for example reassign this customer to a facility that allows for this growth or move another customer to a suitable facility. To onboard a new customer into the 3PL's network, the planning ability of the prescriptive layer should offer an analysis of suitable facilities. It will conduct a feasibility analysis based on storage capacity, throughput capacity and the availability of other necessary services such as the capability to handle a specific type of product.

The prescriptive layer should be able to support more complex capacity planning, encompassing multiple clients over multiple sites. Figure 5 provides a simple yet realistic example of such dynamic planning. On the left side are provided the storage capacity requirement predictions for clients using two facilities in the 3PL's network. In a logistic campus, buildings A and B each currently host three distinct customers. Building A is projected to overflow as capacity requirements from client C are to climb, while building B is projected to be gradually less utilized, mostly related to declining capacity requirements from client F. Smart planning through the prescriptive layer has led to a reshuffling plan with clients D and E shifted to building A, and client C shifted to building B, resulting in smoothing the capacity requirements over the two buildings and avoiding both overflow and underusage, as depicted on the right side of Figure 5.

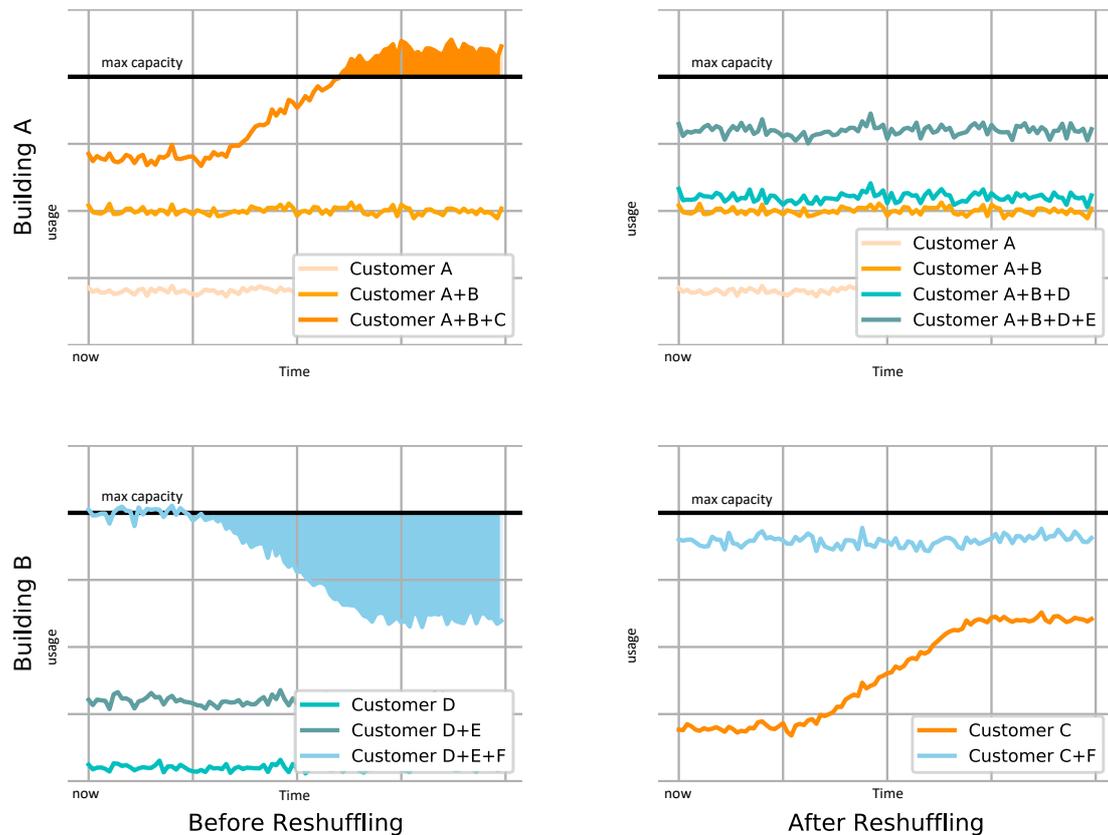


Figure 5: Example of reshuffling suggested in the prescriptive layer addressing storage capacity

The prescriptive layer can to a degree offer decision automation capability, in the line autonomous analytics as proposed by Davenport (2017). Some decisions can be taken by software agent, without direct human intervention beyond setting the agent’s rules and methods, especially those requiring fast response time and taken repetitively over many instances. For most of the higher-impact, more strategic decisions, the prescriptive layer is rather to provide support to human decision-makers. For example software agents can make recommendations and assessing their impact according to multiple metrics, notably through simulation, optimization, and machine learning based methodologies, and then letting the human decision-maker at the 3PL reject, accept, or modify them.

5 Conclusion and Outlook

The three-layer decision-making framework we introduce in this paper for hyperconnected 3PL capacity management allows logistics service providers to counteract the volatility, uncertainty and complexity they are faced with. Through the descriptive layer, the hyperconnected service provider gains insights into the past and current states of the 3PL market, customer activities, contracts, and flow activity in their network. Based on forward looking predictions of these in the predictive layer, the prescriptive layer facilitates decision-making concerning customer and contracting opportunities as well as adapting capacity, assignments and flow within its network. This serves as one thread for a 3PL towards a transformation into a proactive hyperconnected logistics player.

In this work, to the best of our knowledge, we are first to introduce an analytics-based framework for logistics capacity management in the Physical Internet. At each layer of the framework, there is room for future research.

In the descriptive analytics layer, research is notably needed on which information should be shared by logistics service providers and clients in the Physical Internet; how to filter the wide scope and huge scale of information into high-value, focused, and actionable knowledge and insights; how to better leverage novel visual analytics, as well as augmented and virtual reality, technologies; what new key performance indicators should be developed to leverage the hyperconnected essence of the Physical Internet and thus to provide 3PLs with fresh and enlightening perspectives.

In the predictive analytics layer, much research is notably needed on interlacing the various correlated capacity and throughput predictions, to acknowledge alternative probabilistic future scenarios, and to support risk and resilience management in the context of hyperconnected logistic service providers.

Fed by the descriptive and predictive layers, the prescriptive analytics layer opens a wealth of research opportunities for better design and planning of solutions, for better selection between alternative options, for optimizing client, facility, and network wide decisions (e.g. expanding on the example from Figure 5).

From a deeper perspective, the framework allows to break away from rigid contracting modes having been instituted to ensure conservative and robust guidelines and decision framework when having to maneuver a complex organization with minimal timely information availability, minimal predictive capability, and minimal prescriptive decision-support capability. It indeed opens up more hyperconnectivity oriented research and innovation avenues such as considering multiple dynamic external capacity options, and considering smarter and more agile client contracts.

The framework also uncovers relationships between information available to, and decisions taken by, various organizational units within a logistics service provider. This is clearly the case between sales, marketing, and business development; information technology; facilities acquisition, planning and design; transportation and logistics operations. Much research is needed in synergizing these relationships, and guiding decision makers within each unit to take smart decisions with a more holistic perspective.

While the framework is currently being implemented at a major American 3PL player, the initial focus is on putting it into action within a single region, developing the methods, models, and technologies necessary to do so, leveraging cloud technologies. Next efforts are planned to address the whole North American landscape allowing overall visibility and decision-making facilitation on a continental level, and ultimately expanding at a multi-continent international level.

References

- Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, 29(5-6), 594-621.
- Arunachalam, D., Kumar, N., & Kawalek, J. P. (2018). Understanding big data analytics capabilities in supply chain management: Unravelling the issues, challenges and implications for practice. *Transportation Research Part E: Logistics and Transportation Review*, 114, 416-436.

- Banerjee, A., Bandyopadhyay, T., & Acharya, P. (2013). Data analytics: Hyped up aspirations or true potential?. *Vikalpa*, 38(4), 1-12.
- Ballot, É, B. Montreuil, R.D. Meller (2014), The Physical Internet: The Network of Logistics Networks, La Documentation Française, Paris, France, 205p.
- Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. *Journal of the Operational Research Society*, 20(4), 451-468.
- Bennett, N., & Lemoine, G. J. (2014). What a difference a word makes: Understanding threats to performance in a VUCA world. *Business Horizons*, 57(3), 311–317.
- Davenport, T., & Harris, J. (2017). Competing on analytics: Updated, with a new introduction: The new science of winning. *Harvard Business Press*.
- Hahn, G. J., & Packowski, J. (2015). A perspective on applications of in-memory analytics in supply chain management. *Decision Support Systems*, 76, 45–52.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
- Hertz, S., & Alfredsson, M. (2003). Strategic development of third party logistics providers. *Industrial marketing management*, 32(2), 139-149.
- H.R.4040 - Consumer Product Safety Improvement Act of 2008, Sect. 235, 08/14/2008, United States of America
- Huo, B., Ye, Y., & Zhao, X. (2015). The impacts of trust and contracts on opportunism in the 3PL industry: The moderating role of demand uncertainty. *International Journal of Production Economics*, 170, 160-170.
- Marchet, G., Melacini, M., Sassi, C., & Tappia, E. (2017). Assessing efficiency and innovation in the 3PL industry: an empirical analysis. *International Journal of Logistics Research and Applications*, 20(1), 53-72.
- Montreuil, B. (2011). Toward a Physical Internet: meeting the global logistics sustainability grand challenge. *Logistics Research*, 3(2-3), 71-87.
- Montreuil B. (2017). Omnichannel Business-to-Consumer Logistics and Supply Chains: Towards Hyperconnected Networks and Facilities, *Progress in Material Handling Research* Vol. 14, Ed. K. Ellis et al., MHI, Charlotte, NC, USA.
- Montreuil B., R.D. Meller & E. Ballot (2013). Physical Internet Foundations, in *Service Orientation in Holonic and Multi Agent Manufacturing and Robotics*, ed. T. Borangiu, A. Thomas and D. Trentesaux, Springer, p. 151-166.
- Packowski, J. (2013). *LEAN supply chain planning: the new supply chain management paradigm for process industries to master today's VUCA World*. CRC Press.
- Souza, G. C. (2014). Supply chain analytics. *Business Horizons*, 57(5), 595-605.
- Zacharia, Z. G., Sanders, N. R., & Nix, N. W. (2011). The Emerging Role of the Third-Party Logistics Provider (3PL) as an Orchestrator. *Journal of Business Logistics*, 32(1), 40–54.
- Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, 159-175.