European Transport Research Review

ORIGINAL PAPER Open Access

Optimization of a physical internet based supply chain using reinforcement learning

Eszter Puskás[1]*, Ádám Budai[2] and Gábor Bohács[1]

Abstract

Physical Internet based supply chains create open, global logistics systems that enable new types of collaboration among participants. The open system allows the logistical examination of vehicle technology innovations such as the platooning concept. This article explores the multiple platoon collaboration. For the reconfiguration of two platoons a heuristic and a reinforcement learning (RL) based models have been developed. To our knowledge, this work is the first attempt to apply an RL-based decision model to solve the problem of controlling platoon cooperation. Vehicle exchange between platoons is provided by a virtual hub. Depending on the various input parameters, the efficiency of the model was examined through numerical examples in terms of the target function based on the transportation cost. Models using platoon reconfiguration are also compared to the cases where no vehicle exchange is implemented. We have found that a reinforcement learning based model provides a more efficient solution for high incoming vehicle numbers and low dispatch interval, although for low vehicle numbers heuristics model performs better.

Keywords: Physical internet, Supply chain, Virtual hub, Platoon, Reinforcement learning

1 Introduction

In recent years, achieving sustainable operations has been a major driving force both in logistics and also in the automotive industry. From the future concepts of logistics systems one of the most outstanding ideas is the Physical Internet (PI), which was conceived by Montreuil [1]. The idea debunks the regular methods and practices of transport, warehousing and material handling. The PI establishes a completely new structure for the operation and logistics networks. Similarly to the flow of information based on the Digital Internet data packet, goods flow through the network in a specially designed container (π container), which has all the features needed for sustainability and efficient operation. The concept is based on network-level collaboration to create an "open global logistics system". Because of the novelty of the Physical Internet concept, a breakthrough is yet to come, but the most innovative companies are already testing the model and pilot projects have been implemented, such as MonarchFx, Carrycut or CRCServices [2].

In addition to the development of the logistics area, significant innovations have emerged in the automotive industry as well. The real revolution in freight transport is the concept of platooning, originally proposed by California Partners for Advanced Transportation Technology (PATH) [3]. Platooning represents a set of vehicles on the road in which the distance between the neighboring vehicles are significantly smaller than human drivers can maintain without risk [4]. The short distance is achieved through continuous communication via V2V communication. As a result of the short distance between trucks, we can increase total road efficiency [5] and improve the aerodynamics of all trucks, thereby reducing fuel consumption citePatten, Alam. Conceptually, a platoon consists of a leader (first in the line) and one or more follower

*Correspondence: eszter.puskas@logisztika.bme.hu
[1] Budapest University of Technology and Economics, Faculty of Transportation Engineering and Vehicle Engineering, Dept. of Material Handling and Logistics Systems, 1111 Bertalan L. u. 7-9., Building L., Budapest, Hungary
Full list of author information is available at the end of the article

(all others in the line) vehicles [6]. Once the vehicles are connected, the driver must sit in the first vehicle while all other vehicles follow autonomously the leader's activity and speed. The great interest in the concept is due to its large potential for reducing transport costs by reducing fuel consumption. According to past data, 95% of accidents are caused by people [7]. At the social level increasing safety through driving automation is important [8]. Finally, the concept reduces congestion and traffic jams, increases freeway utilization, while reducing greenhouse gas emissions and air pollution [6].

Janssen et al. (2015) investigated the possible effects of platooning on the heavy-duty vehicles (HDVs) supply chain processes [9]. The benefits will be greater if they can cooperate with other suppliers, even competitors [6].The Physical Internet concept for the logistics network provides the environment for an efficient platooning system. In an open global supply chain a common platform and language is provided for seamless communication between the HDVs [10].

The future logistics challenges are sustainability, low emissions and resource efficiency. In our opinion, from the logistical point of view, the most important task is the optimal operation of the system's components. For this, it is particularly important to define the exact operating model [11].

This article is about formulating a new model for better utilization of our existing resources, taking into account vehicle technology trends and the application of new principles based on the Physical Internet. The rest of this article is structured as follows. Section 2 discusses relevant literature studies. Sections 3.1 and 3.2 details the applicability and limitations of the proposed model. The purpose during optimization is to minimize fuel costs, waiting costs and labour costs. Section 3.3 describes the heuristic model and the reinforcement learning based model for controlling platoon reconfiguration. Section 4 presents a numerical example comparing the basic platoon model to the heuristic and the reinforcement learning method. Section 4 also discusses the main results of the model's analyses. Section 5 summarizes the results and points out future research.

2 Literature review

Based on the statistics of recent years, carbon dioxide emissions continue to rise. Reducing this is considered to be one of the most difficult challenges of the economy, given the continuing demand for road freight transport [12]. Last year, the European Commission adopted an EU standard for CO2 emissions from heavy-duty vehicles which says that by 2025 average CO2 emissions should be 15% lower [13]. To achieve sustainability, resource efficiency and low emissions, the platooning concept has recently gained increasing interest, mainly in terms of technical, safety and autonomous management [6]. Some projects implement and demonstrate these aspects, for example PATH, SARTRE, CHAUFFEUR or KONVOI [14].

Applying the platoon concept will change the supply chain network, leading to system-wide innovation. The study of The European Truck Platooning Challenge 2016 confirmed that creating, managing and optimizing platoons is a major challenge [9, 15]. To exploit the benefits of the platooning system innovative decision models are needed. In addition to methodological improvements, they also help to quantify the benefits [16].

To take advantage of the platooning concept, it may be advisable to deviate slightly or more from the optimal route between the starting point and the destination. In addition to solving the detour problem, the decision support model operation is further complicated by the lead-follower decision situation of the platoon concept [14]. Van De Hoef et al. (2015) examined the case of two vehicles with fixed paths intersecting each other. By setting the correct speed, the two vehicles reach the designated intersection at the same time [17].

The model developed by Wei Zhang et al. (2017) provides useful insights into the effect of travel time uncertainty. In their article it had been shown that if the scheduled arrival times of the vehicles are different, the goals of saving fuel and arriving on time were conflicting [18]. A similar issue of energy-delay trade-off has been analyzed in [19]. The authors presented an algorithm that highlights the negatives of the platoon concept. Even though vehicles that are in a platoon can reduce energy consumption and emissions, the accumulated waiting time for the interconnection can greatly increase the delay [19].

The hub-based network offers an opportunity to explore the platooning concept from other aspects. In this case, the vehicles spend their entire path between the two hubs. Rune Larsen et al. in their article extended the model by [20] by introducing two different methods. In the case of fixed hubs and vehicles with fixed paths, coordination of platooning creation has been optimized, as illustrated in Fig. 1. A virtual hub centre and a collaborative platooning system was assumed. In their model, they assumed full cooperation between participants, regardless of manufacturer or supplier. The importance of collaboration is highlighted by [21], who demonstrated by analysis that it is hard to effectively create platoon groups when a significant number of trucks are unable to connect or because of physical limitations, e.g. the maximum length of the platoon is low or there are strict time windows for shipments [21]. The estimated profit is about 4-5% of the fuel or about 38 €per trip for long-distance trucks. In a competitive market this profit can be a key issue [22].

Saad Alshiddi et al. presented the possibility of connecting two different platoons. In the first case, one of

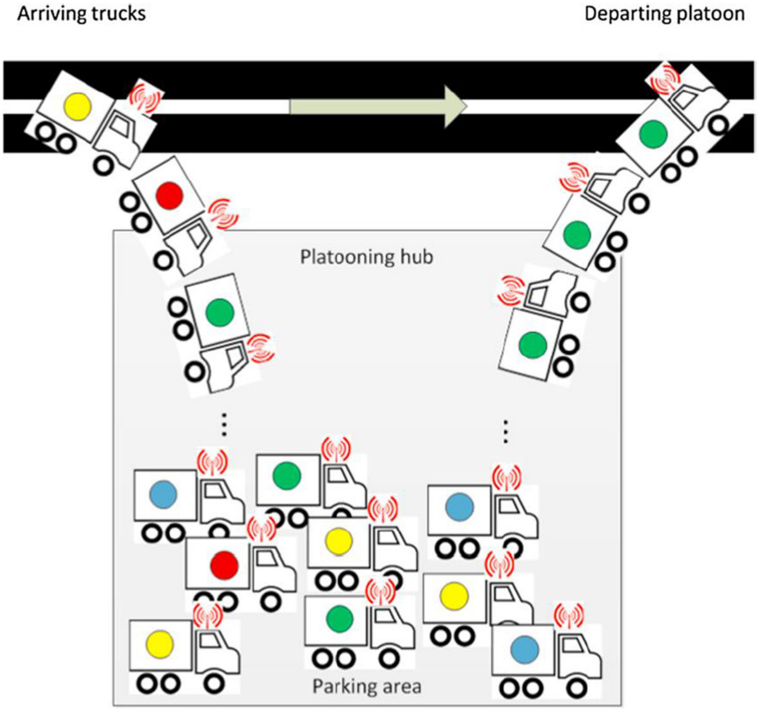

Fig. 1 Platooning hub [22]

the platoons will check at the moment of departure to determine if there is any existing platoon with the same destination that you want to join. The connection is possible if the maximum number of vehicles and the maximum waiting time are not exceeded. In the other case, both platoons are on their way and are searching with GPS possible platoons with the same destination to which they could connect [23].

We can see that several papers address the issue of platoon organization and all studies acknowledge that platoon planning problems are difficult to solve. Most studies focus on the platoon created by the collaboration of different companies. Fewer studies can be found on the design of collaborations between several platoons launched from different locations. In our opinion, one of the interesting directions of future research is to examine the cooperation of several platoons. There is a lack of research to define the problems generated by collaboration and to compare different solution approaches. Another interesting direction could be to examine the effect of various constraints and parameters on operation.

In this article we would like to further increase the benefits of platooning in a Physical Internet based logistics network. According to the concept outlined in [22], the platoons in the examined system travel only between hub centres. To take advantage of the connection we aim to create a model in which platoons can reconfigure themselves. The platoons are launched between hubs at fixed intervals. The selected platoons will come across at a virtual hub point where the platoons can change their vehicles based on their destination. The model aims to reduce the fuel cost, labour cost and cost of the waiting time defined by the objective function.

3 Methods

In this paper, we create a new theoretical model to better exploit the benefits of a platoon system. In the study the environment is Physical Internet based. We assume that all HDVs interact and communicate with each other and the hubs are open and accessible to all PI users. The presented model extends the transportation model outlined in the article [24]. As stated in the article [24], the objective is to optimize the shipping tasks performed by platoons between predefined hubs. This requires a virtual hub to provide a common meeting point for the selected platoon. Beyond determining the location of the virtual hub, we also examine the methodology for selecting the platoons and vehicles to be paired.

The block diagram of the model is shown in Fig. 2. The figure shows the required inputs and the outputs. The diagram shows the two methodologies used to solve the platoon cooperation: the heuristic-based algorithm and the reinforcement learning-based algorithm. Furthermore the figure represents the assumptions, such as the capacity limit for the platoons.

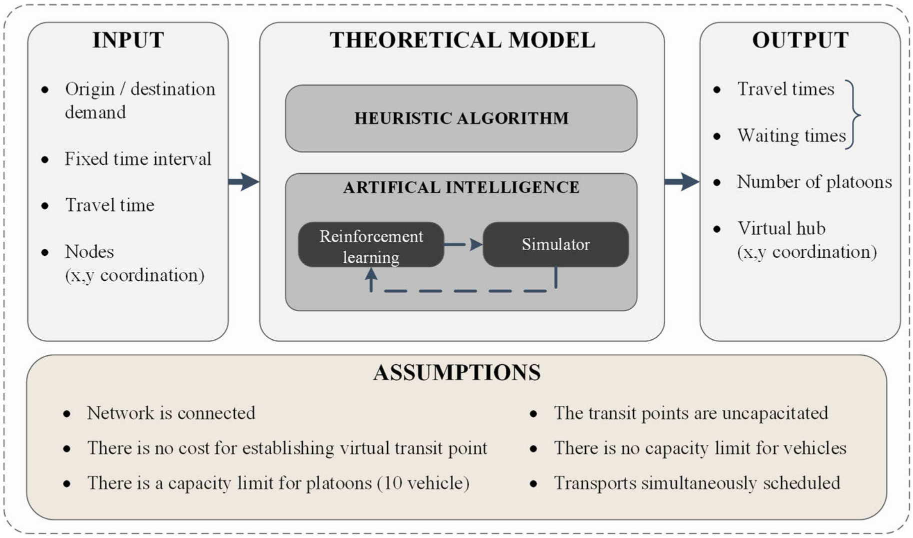
INPUT
Origin / destination demand
Fixed time interval
Travel time
Nodes (x,y coordination)
THEORETICAL MODEL
HEURISTIC ALGORITHM
ARTIFICAL INTELLIGENCE
Reinforcement learning
Simulator
OUTPUT
Travel times
Waiting times
Number of platoons
Virtual hub (x,y coordination)
ASSUMPTIONS
Network is connected
There is no cost for establishing virtual transit point
There is a capacity limit for platoons (10 vehicle)
The transit points are uncapacitated
There is no capacity limit for vehicles
Transports simultaneously scheduled

Fig. 2 Theoretical model. Theoretical model for virtual transfer point optimization in Physical Internet network (based on [24])

The inputs, outputs, limiting conditions, the target function of the model and the solution methodologies are detailed in the following subsections.

3.1 Limitations and inputs

Similarly to the article written by [22], any restrictions that may result from technological sources and legislation will be removed. In addition, our simulation does not handle taxes and tolls. From a network point of view, the assumptions can be very restrictive but are relieved in this article. Based on the Physical Internet system, the network is completely interconnected and the flow of information between all participants is ensured. There are no restrictions on the virtual hub, so there is no cost for establishing and there is no capacity limit. The simulation does not take into account the capacity of the vehicle, since the basic object in the simulation is the vehicle itself, not the products to be transported by the vehicle. In the developed simulation each platoon can consist of a maximum of 10 vehicles [25]. From fixed hub centres platoons are created at fixed intervals. Launch frequency is an important variable for strategy which we will test across a range of values.

The first input of the simulation is the departure and destination coordinates. These stations do not change during the simulation. The simulation generates the incoming vehicles and the associated basic parameters. This information is the departure and destination of the vehicle and the time and distribution of its arrival in the system. As mentioned, the platoons are launched at fixed time intervals. The simulation does not include travel uncertainty.

3.2 Problem formulation

In this article, we included the minimization of fuel consumption costs among the goals of the method. Fuel consumption depends primarily on the traveled distance. The fuel consumption of HDV has decreased significantly in recent years. From 41.9l/100km, measured in 2002, the consumption was reduced to 35.6l/100 km [26]. Delgado et al. tested various vehicles under three types of load in urban, regional and long-distance transport. Among the values they examined, the average load consumption of HDV was 36.4l/100km for regional transport and 33.1l/100 km for long-distance transport [12]. Based on the publications presented, our model assumes a fuel consumption of 35l/100km and will use a fuel price of 1.2€/l by[22] article. One of the strongest economic benefits of the platooning concept is that vehicles can save fuel. Fuel savings from platooning have been investigated in numerous projects. Pilot studies show that fuel consumption can be reduced by up to 15% [27]. In the article written by Lammert et al. based on the presented model their result was 5.3% saving on the leading vehicle. In the case of vehicles following a significant fuel saving of 9.7% was measured [28]. In this article, we assume a 10% savings on the following vehicles, while the leader vehicle does not count on any fuel savings.

When creating a platoon, vehicles are forced to compromise to gain common benefits. Such a trade-off is, for example, the cost of waiting time when the two vehicles do not arrive at the target virtual hub at the same time. Sokolov et al. investigated the effect of waiting time on the benefits of the platooning concept and they have found that the waiting cost was 27.17€/hour [29]. Larsen et al.

calculated with a 35euro/hour waiting cost, the salary of the most expensive drivers [22]. In Zhank Wei's article the cost of waiting is divided into two types of charges: the penalty for the early arrival of the vehicle is 0.0093€/sec and for delay is 0.0466€/sec [30]. Based on these, our article assumes a 35€/hour waiting cost.

In this paper, each platoon's leading vehicle also has a driver, while the following vehicles are autonomous. The concept justifies taking into account the labour cost generated by drivers, which is 0.73€/km based on [31, 32].

We consider the problem in the setting that a set of homogeneous vehicles {1, 2, ..., V} enter the system with the same priority. Each vehicle is assigned transport tasks that define departure and destination. There is no time window set to the vehicles, so there are no costs associated with delays or earlier arrivals. It is assumed that the vehicles move at the same and constant speed in the network.

Using the generic notations compiled in Table 1, the total cost for the model consists of the following objective function (1) , which is further explained in (2) - (5).The restrictions are in (6) - (9).

Table 1 Mathematical notation

Parameter	**Description**
i	Set of input nodes
j	Set of outputs
v	Set of leading vehicles of platoons
w	Set of vehicles
p	Set of platoons
V_i	The amount of trucks arriving at i starting nodes during the entire time horizon
M	Maximum number of truck in each platoon
P	Set of platoons
$s(i,j)$	Distance traveled by the vehicle between points (i,j)
t_{in}	Arrival time of truck v_i
t_{disp}	Dispatched time of truck v_i at input nodes
t_{hub}	Waiting time of vehicle v_i at the virtual hub
c_w	The cost of waiting an hour
c_p	Driver labour cost per kilometer
c_f	Unit cost of fuel
β	Platoon follower indicator
η	The amount of fuel cost reduction
Variable	**Description**
$x_{v,p}$	Is equal to 1 if and only if truck $v \in V$ is assigned to platoon $p \in P$
$x_{v,w,p}$	Is equal to 1 if and only if truck $w \in V$ follow truck $v \in V$ in platoon $p \in P$

The total cost to be minimized consists of three components: fuel cost, waiting cost and labour cost. All three kinds of costs are modeled as linear functions. Based on the equations detailed below, we summarize the three components converted into monetary units (EUR). The total cost can be expressed as:

$$C(\Theta) = C_f(\Theta) + C_t(\Theta) + C_p(\Theta) \tag{1}$$

where $C_f(\Theta)$ represents the fuel cost, $C_t(\Theta)$ is the cost of the waiting, and $C_p(\Theta)$ represents the driver labour cost of the platoon leader and Θ represents a set of decision variables, for example pairing choice.

Fuel cost is calculated using the equation defined in [30] on the traversed edge (i,j):

$$C_f = \sum_{V_i} c_f s(i,j)(1 + \beta(\eta - 1)) \tag{2}$$

$$\beta = \begin{cases} 1 & \text{if the vehicle is a platoon follower} \\ 0 & \text{if the vehicle is a platoon leader} \end{cases} \tag{3}$$

From Eq. 2, the fuel cost is the product of the fuel unit cost (c_f) and the distance (s(i, j)) travelled by the vehicle between points (i, j), given that the vehicle is a leader or follower in the platoon. The s(i, j) depends on the routing. Platoon cooperation defines the distance that vehicles have to travel. If one platoon is to cooperate with another platoon, the vehicles must take a detour to touch the meeting point, which is a virtual hub. The resulting product is calculated for each vehicle over an 8-h time horizon and then summed to give the total fuel cost. According to Eq. 3, if $\beta = 0$ the vehicle is a leader of the platoon and for $\beta = 1$ the vehicle is the follower. In addition, the parameter $0 < \eta < 1$ determines the amount of fuel cost reduction. Assuming a 10% savings based on literature research, we use $\eta = 0.9$. This means that if the vehicle is a leader, it does not mean any savings for it to drive as a platoon, whereas in the case of a follower vehicle, we can expect a 10% reduction in fuel.

As mentioned in the previous paragraph, the time horizon represented in the simulation is 8 h. It starts from the zero minute and lasts until the 480 min. The vehicle could wait in two locations. First at the starting station, as the simulation dispatched the platoons at fixed intervals. To achieve synchronization at the virtual hub, one platoon can wait for the other, generating a waiting time in the system. The second vehicle waits at the virtual hub, where one platoon can wait for the other platoon due to the different paths of the two platoons. The vehicles arrive at the departure stations at different times. Until they are launched, they wait at the starting station within the time horizon. Based on these, the waiting time is calculated as follows:

$$C_t = \sum_{V_i} ((t_{disp} - t_{in}) + t_{hub}) c_w \tag{4}$$

The waiting time cost for each vehicle is calculated by multiplying the waiting time per unit cost of waiting time. The $(t_{disp} - t_{in})$ subtraction represents the time spent at the starting node while the second part t_{hub} represents the time spent at the virtual hub.

The driver labour cost of the platoon leader is calculated as follows:

$$C_p = \sum_{V_i} c_p s(i,j)(1+\beta) \tag{5}$$

We can determine the number of platoons launched within the 8-h time horizon examined. For each platoon, a predetermined vehicle is the leading vehicle of the platoon, which is represented by the β parameter shown above. If the vehicle is a platoon leader, the labour cost is the product of the distance travelled and the driver labour cost per kilometer.

$$\sum_{p \in P} x_{v,p} = 1 \tag{6}$$

$$x_{v,p} \in \{0;1\} \quad \forall\, v \in V, p \in P \tag{7}$$

$$x_{v,w,p} \in \{0;1\} \quad \forall\, v, w \in V, p \in P \tag{8}$$

$$\sum_{v \in V} x_{v,p} \leq M \quad \forall\, p \in P \tag{9}$$

Constraint conditions are also linear functions of variables. Equation (6) ensures a truck is allocated to exactly one platoon. Equations (7) and (8) restrict the domain of the decision variables. Equation (9) restricts the size of platoons to be less or equal to M.

The goal is to find the best possible platoon pairing so that the resulting transportation tasks minimize the total cost defined in the objective function (1). A possible solution for pairing using the heuristic-based algorithm is shown in 3.3.1 section. Then, to minimize the objective function, we present a RL algorithm in 3.3.2 chapter.

3.3 Solution methodology

In recent years, both professionals and academics have devoted a great deal of attention to study the supply chain [33]. Although artificial intelligence (AI) seems to have been promising in the development of human decision-making processes since the 1970s, it has been used to a limited extent in the supply chain [34]. Despite the challenges, ongoing research into AI is a promising area in the supply chain [34]. Among the AI methodologies, the reinforcement learning (RL) technique is receiving increasing attention. The widespread use of the RL methodology in logistics dates back to 2002. Pontrandolfo et al. (2002) studied the problems of global supply chain management using the RL technique [33]. Stockeim et al. (2002) used the RL technique to solve a decentralized supply chain problem, where he inserts the stochastic demands into the production queue in search of an optimum [35]. Transport is an important driver of the supply chain as products are rarely manufactured and consumed in the same field. Habib et al. (2016) explored the possibility of RL based route optimization [36]. A key aspect of the methodology is simulation which subserves the widespread application of the methodology [33]. To the best of our knowledge, there is currently no method for managing platoons in a Physical Internet based logistics network that utilizes RL. Therefore, to our knowledge, this work is the first attempt to apply an RL-based decision model to solve the problem of controlling platoon cooperation.

In the next subsections, two different solution methodologies are proposed. Both solutions are responsible for selecting the nodes and vehicles for platoon pairing. The first algorithm provides a heuristic based solution. The second algorithm uses deep reinforcement learning to minimize the objective function. This second algorithm provides a much more general solution methodology that can be applied to larger and more complex networks as opposed to the first algorithm. This article does not provide mathematical guarantees for optimal or even locally optimal solutions.

3.3.1 *Algorithm - Heuristic*

In the first solution, the goal is to provide a solution for platoon collaboration. The algorithm is illustrated in the flowchart shown in figure 3.The input values include the coordinates defining the origins and destinations, as well as the fixed time interval, which determines at which intervals the platoons will dispatch. Finally, the arrival time of the vehicles at the origins is also input data. The first step of the algorithm is to calculate the sum of vehicles per origin-destination pair at the set time intervals. For example, after 30 min (if this is the set fixed time), how many vehicles are waiting at "origin-1" whose transportation task is to reach "destination-2". This creates a table in which each row corresponds to the total number of vehicles in an origin-destination pair in descending order. From this table, we select the first two rows, so the origin-destination pair with the two largest vehicle numbers, thereby defining the directions involved in the collaboration.

The next step is to select the vehicles to be dispatched from those waiting in the specified starting directions. For that, we create a new table that includes the vehicles whose transport task is the previously selected origin-destination pair. For selection, we arranged these vehicles in descending order according to the waiting time at the starting station. Based on the table, we select the vehicles

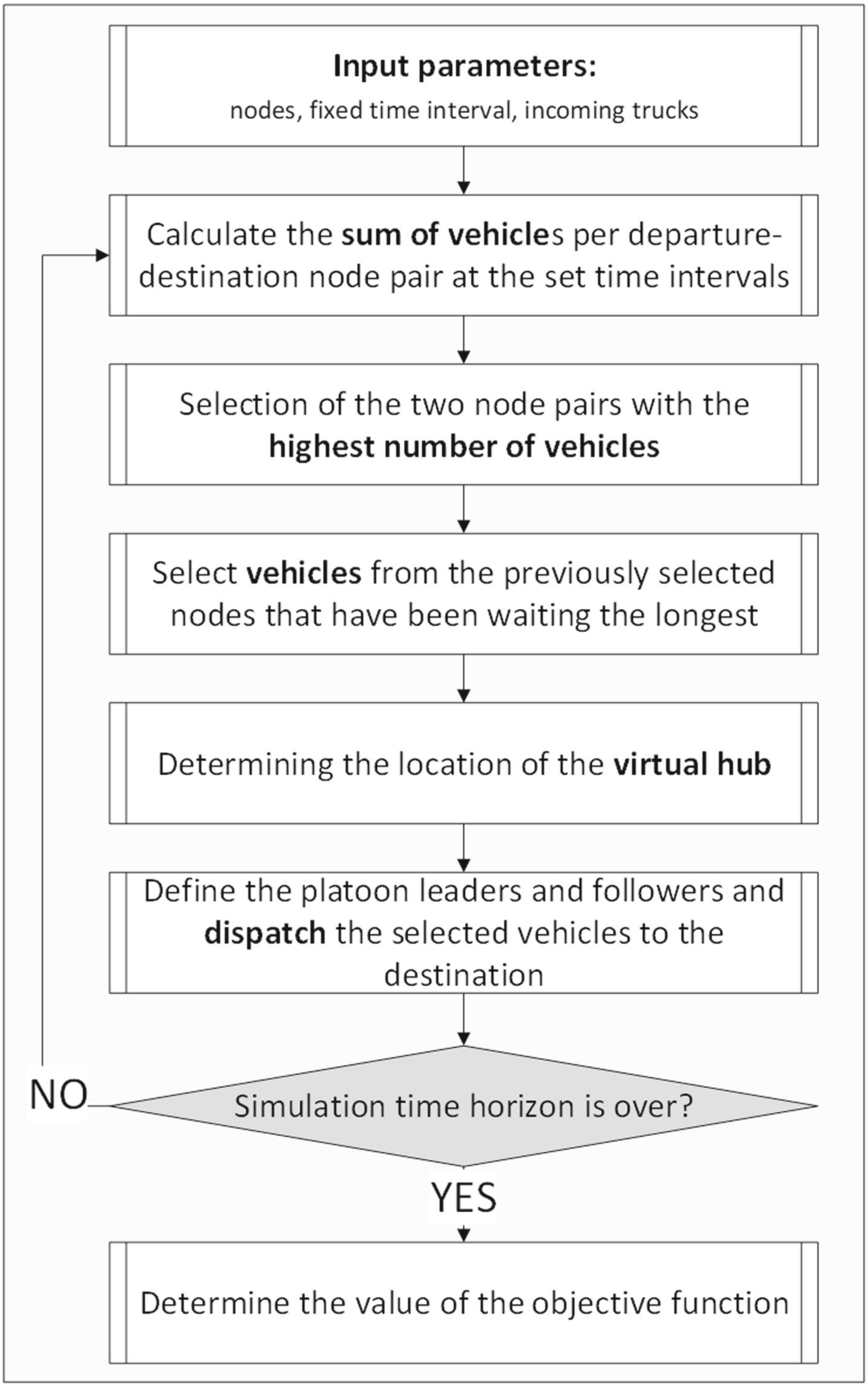
Input parameters:
nodes, fixed time interval, incoming trucks
Calculate the sum of vehicles per departure-destination node pair at the set time intervals
Selection of the two node pairs with the highest number of vehicles
Select vehicles from the previously selected nodes that have been waiting the longest
Determining the location of the virtual hub
Define the platoon leaders and followers and dispatch the selected vehicles to the destination
Simulation time horizon is over?
NO
YES
Determine the value of the objective function

Fig. 3 Flowchart of a first algorithm - Heuristic

that have been waiting for the longest time, taking into account that there can be a maximum of 10 vehicles in a platoon during the entire transport based on the restrictive condition. At the specified fixed intervals, only one platoon can be dispatched from each origin at the same time.

For cooperating platoons, we need to determine the location of the virtual hub needed for reconfiguration.

The virtual hub will be the place where cooperating platoons meet and vehicles can change platoon. Its location is determined by the centre of gravity method. The two origins and the two destinations define a rectangle. The weight of the rectangle vertices is the number of vehicles dispatching from the origins and arriving at the destinations. From the non-selected origins, all waiting vehicles (up to 10 vehicles due to the restrictive condition) form a platoon from each direction, performing the transport task without cooperation. This is called the basic model. If the vehicles in the platoon have different destination purposes, the platoon must reach each of them. The platoon leader has to be a vehicle which destination is the last destination the platoon visits.

The algorithm repeats these steps in each fixed time interval until the simulation time horizon ends. At the end of the simulation, the value of the objective function is determined. This algorithm provides a solution for platoon collaboration but does not provide mathematical guarantees for optimal solutions.

3.3.2 Algorithm - reinforcement learning

Nowadays, deep reinforcement learning enjoys high popularity due to the breakthroughs experienced in the last decade. A series of papers have proven that combining deep neural networks with reinforcement learning can be trainable and beneficial [37, 38]. The goal of reinforcement learning is to train an agent to behave optimally in a given environment [39, 40]. Figure 4 shows the main components of reinforcement learning from a high-level perspective. The agent interacts with the environment by actions. The agent can observe the state of the environment and the environment signals a reward to indicate the quality of the action. In this paper, the agent represents the dispatcher and makes decisions about the platoon collaboration of each vehicle. That is, based on the flowchart showing the previous algorithm (see Fig. 3), the RL-based algorithm replaces the second and third steps of the figure. In the other steps, the two algorithms are the same. The environment is the collection of vehicles, starting points and the destinations. The reward is the negative of the total cost. This consists of the fuel cost, waiting time cost and labour cost.

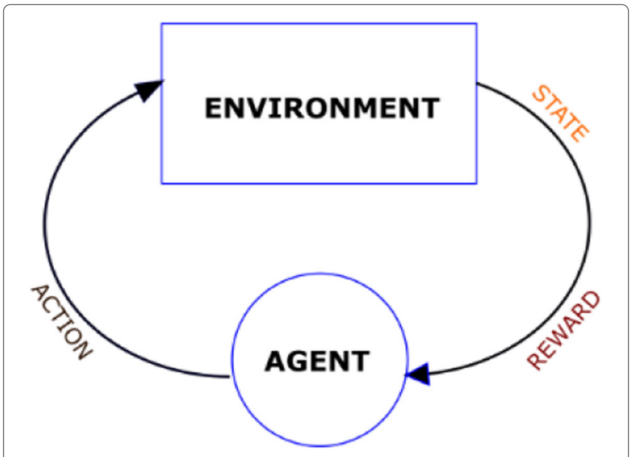

Fig. 4 Reinforcement learning framework

The state describes the environment in time. In this case the state is represented as a 60 times 3 sized matrix. There are 20, 20 and 20 vehicles for each destination. The first column gives the time the vehicle waited so far at the origin. The source (or origin) is given by the second columns (values can be 1, 2, 3) and the last column shows which vehicle is chosen to depart in the current cycle. Therefore the size of the state space (if we just look at column 2 and 3) is greater than 6^{60} and this fact motivates the usage of deep reinforcement learning. The agent's action changes the state of the environment according to its dynamics which is stochastic in this case. Now, the action is represented as an integer number from [0, 9). Each number encodes a source-destination pair. The algorithm chooses source-destination pair. Vehicles departing from the source and heading to the destination of the chosen pairs are not involved in platoon collaboration. They can simply avoid the hub, see Fig. 5. In reinforcement learning, the goal of the agent is to maximize the expected return. The return is the accumulated reward during the whole process. The following formula defines the return (G):

$$G = \sum_{t=0}^{T} \gamma^t R_t \tag{10}$$

Where γ is the discounting factor and its value can be between 0 and 1, exclusively. The decision-making mechanism of the agent is modeled with a function, the so-called policy. This function is a mapping between the states and actions. Then, reinforcement learning can be formalized in the following way:

$$\pi^* = \arg\max_{\pi} E_\pi [G] \tag{11}$$

Where π is the policy. In order to find the optimal policy three main approaches were developed. Value-based [41, 42], policy-based [43] and actor-critic methods [44]. In this paper we utilize the DQN algorithm [41]. This algorithm is based on the action-value function (Q):

$$Q^\pi(s, a) = E_\tau [G(\tau)|s_0 = s, a_o = a, \pi] \tag{12}$$

Where τ means the trajectory, s stands for state and a for action. The trajectory is the sequence of states, actions and rewards in time. The action-value function shows the

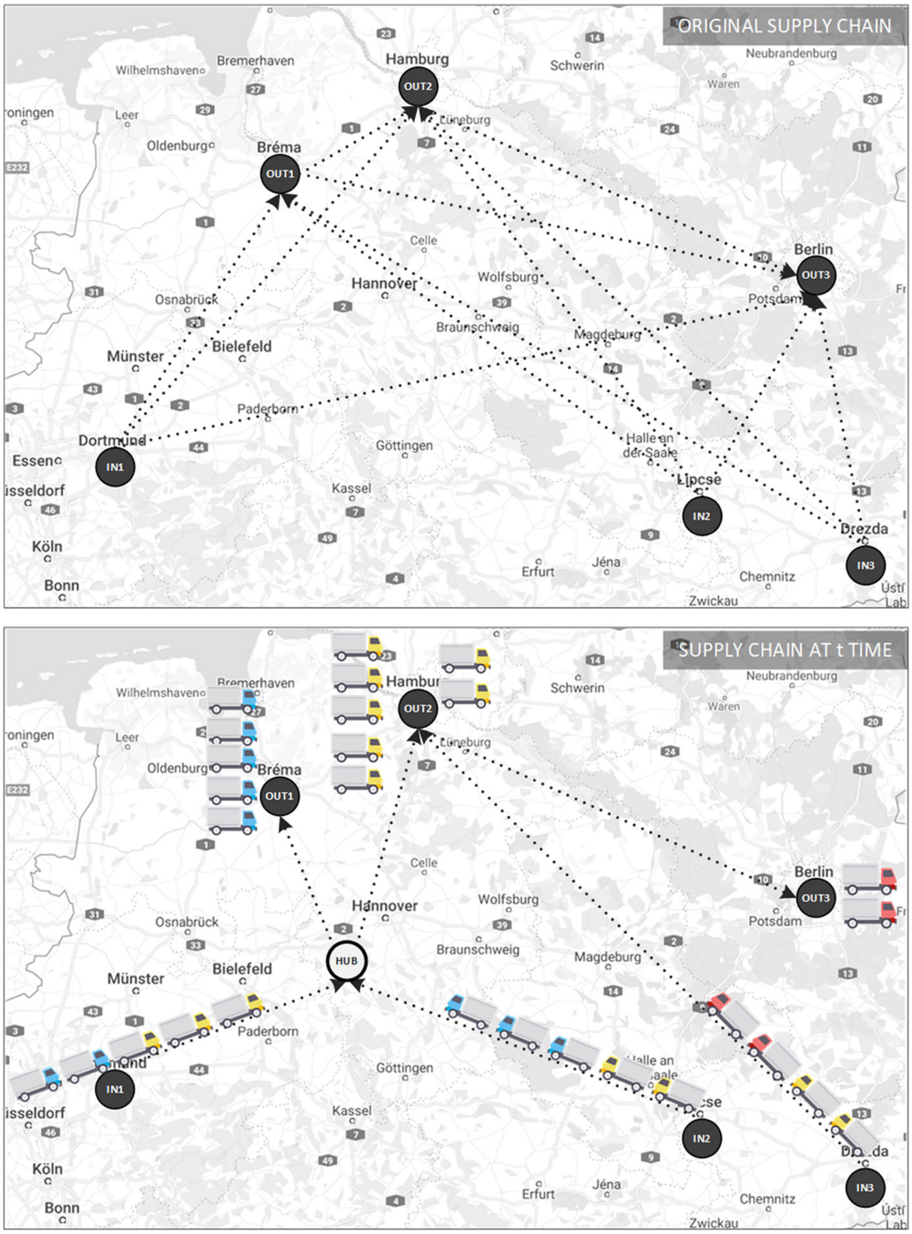

Fig. 5 The structure of the example network

value of each action in each state. Therefore, by knowing the optimal Q-function, the policy can be derived as:

$$\pi^*(s) = \arg\max_a \left[Q^*(s,a) \,|\pi \right] \quad (13)$$

The representation of the Q-function tends to be challenging when the number of possible state and action pairs are very high. This happens frequently and it is known as the curse of dimension. This can be tackled with applying neural networks to represent the Q-function. Our network architecture contains two fully connected layers with 16 and 9 units. The first layer has Rectified Linear Unit (ReLU) activation while the second has a linear one. The reward function was formalized based on Section 3.2.

The training of the Q-network uses the following update rule [41]:

$$\Theta_{t+1}^{upd} = \Theta_t^{upd} + \alpha \cdot \left(r_t + \gamma \max_{a'} Q_{\Theta_t^{frz}}(s_{t+1}, a') - Q_{\Theta_t^{upd}}(s_t, a_t) \right) \frac{\partial Q_{\Theta^{upd}}(s_t, a_t)}{\partial \Theta^{upd}} \quad (14)$$

Where the Θ^{frz} is the weight of the frozen network and Θ^{upd} is the network updated in each iteration. The application of the two networks makes the training more stable because the supervised signal $(r_t + \gamma Q_{\Theta_t^{frz}})$ remains similar for several training cycles. The frozen network is the delayed version of the network with Θ^{upd} parameters. But the frozen network has to slowly follow up the changes in the updated network before it becomes outdated. Therefore the two networks (same architecture but they differ in parameters) are synchronized according to soft update, see Eq. 16. During training we utilized a Boltzmann-sampling for choosing the next action, see Eq. 16. Boltzmann-sampling chooses the next action according to the action values. Therefore actions with similar values are taken into account with similar probabilities, providing the chance to decide which one is really better. The Boltzmann-sampling ensures a balance between exploration-exploitation and helps to discover the environment at the very beginning and conclude in the optimal policy at the end. The DQN algorithm uses several hyper-parameters, for the summary see Table 2. We found this parameters to perform best after an extensive grid search and we experienced that the algorithm is quite robust around these parameters.

$$\pi(s,a) = \frac{e^{-Q(s,a')\tau}}{\sum_{a'} e^{-Q(s,a')\tau}} \tag{15}$$

$$\Theta_{t+1}^{frozen} = (1-\varepsilon)\Theta_t^{frozen} + \varepsilon\Theta_t^{updated} \tag{16}$$

We implemented a simulator for the environment which is gym compatible [45]. The source code is available on github [46]. The logic of selecting vehicles is the same as the one used in the heuristics. For non-collaborative (non-selected) directions, the simulator works the same as for a heuristic-based model.

4 Results and discussion

The presented novel models for the reconfiguration of the platoons are tested via a numerical example and this section presents the simulation results. The numerical example is illustrated by the simple supply chain shown in the Fig. 5.

In the example, three departure stations and three destinations were assumed. According to the map, the departure stations are Dortmund, Leipzig and Dresden, and the destinations are Bremen, Hamburg and Berlin. The simulation results are obtained using our simulator written by Python. The calculation of distances between cities was provided by the Python Geopy module. The results correspond to the mathematical models presented in Section 3. We consider the speed of every vehicle: v = 70km/h. Vehicles are generated from a Gaussian distribution with different parameters for every from-to pair. Table 3 shows the parameters for the Gaussian distribution used for each departure and arrival relation. The simulation was performed for four different generated inputs which differ in the mean value used for the Gaussian distribution. The mean values for each run are shown in the last four columns of Table 3.

The vehicles generated with four different parameters were compared to the results of simulations based on three different methods. These are the basic model, the heuristic algorithm and the RL algorithm. In the case of the basic model, the required destinations will be visited in the order of Bremen - Hamburg - Berlin on a round trip. The simulation model for heuristics and deep reinforcement learning is implemented as outlined in Section 3. For each of the four vehicle generations with different parameters, a simulation test was performed for a total of 11 different fixed review times: from a 15 min review time to a 60 min review time using 5-min increments. We ran the simulation 10 times for 480 min in each case, that is, a combination of 11 different fixed time intervals and 4 generated sets of vehicles. Thus, a total of 440 simulations were performed.

During the simulation runs we determined the total cost of the run, which is presented in Section 3, and the

Table 2 Hyper-parameters of the RL algorithm

Hyper-parameter name	Value
Batch size	32
Discounting factor	0.99
Optimizer	Adam
Learning-rate	5e-4
Synchronization freq. (ε)	1e-2 (16)
Experience replay mem. size	1000
Policy for exploration	Boltzmann-sampl. (τ = 1.0) eq. 15
Training length	10000

Table 3 Parameters for the Gaussian distribution

From	To	Deviation	Mean-1	Mean-2	Mean-3	Mean-4
1	1	2 min	2 min	4 min	8 min	16 min
1	2	4 min	3 min	6 min	12 min	24 min
1	3	1 min	4 min	8 min	16 min	32 min
2	1	2 min	5 min	10 min	20 min	40 min
2	2	5 min	1 min	2 min	4 min	8 min
2	3	3 min	4 min	8 min	16 min	32 min
3	1	1 min	6 min	12 min	24 min	48 min
3	2	2 min	4 min	8 min	16 min	32 min
3	3	2 min	3 min	6 min	12 min	24 min
Average of the mean values (λ)			3,56 min	7,11 min	14,22 min	28,44 min

number of vehicles launched during the run. Because the number of vehicles launched during different runs may vary, the comparison was made based on the total cost per vehicle. The cost per vehicle is the quotient of the total cost and the number of vehicles.

As a first step in the analysis, a comparison was made for vehicles with an average arrival time of 3.56 min (according to Table 3), using the results obtained for all fixed intervals. Accordingly, we compared the heuristic-based algorithm and the reinforcement learning algorithm. The results are shown in Fig. 6.

The results show in this case that reinforcement learning performs better than heuristics. On average, a lower total cost can be achieved by using the reinforcement learning method instead of the simple heuristics if vehicles arrive frequently. We will examine this in more detail below, such as how the results change with less frequent arrivals.

Figure 7 shows simulation results for the three models (basic, heuristic, RL) with an average arrival time of 3.56 min which is a more detailed version of the previous Fig. 6. Figure 8 shows an average arrival time of 7.11 min. The graphs show the total cost per vehicle as a function of the different fixed time interval at which vehicles in the system are dispatched according to the appropriate method.

The two diagrams show that in these two cases the RL-based methodology performs better than heuristics. The basic model, when the platoons do not cooperate with each other after a specified time interval, was operating at a lower cost. With an average arrival time of 3.56 min after a 25 min review time, and with an average arrival time of 7.11 min after a time interval of 40 min. This can be explained by the geographical location of the cities in the network used in the example and the high number of vehicles grouped in a platoon.

Figures 9 and 10 show, respectively, the results of the three methods compared with an average arrival time of 14.22 for Fig. 9 and with an average arrival time of 28.44 for Fig. 10.

In the case of the diagrams in Figs. 9 and 10 the heuristics method performs better, so in these two cases the presented RL-based algorithm (unlike the results in the Figs. 7 and 8) does not perform better. This is due to the low number of vehicles and the high arrival time. Thus, if vehicles arrive infrequently, few vehicles will flow in the system, resulting in less data. Owing to the small number of vehicles, the RL algorithm did not sufficiently learn how to operate the model. Learning cannot be improved in this case since there is too little data available and the time horizon is fixed. The basic model performs better than pairing if you increase the fixed time interval and still have enough vehicles to reach the maximum platoon size allowed.

Figure 11 illustrates the change in platoon size. The figure shows the average vehicle number for the four different input data generated. By increasing the fixed time interval at the departure, we can see that we can create longer platoons. We reach the maximum platoon number sooner with the more vehicles we have available, so the lower the average arrival time (λ) of the generated vehicles is.

Overall platoon cooperation is more profitable with shorter time intervals and higher incoming vehicle numbers. If we can create platoons including 10 vehicles, the basic model-based operation will result in lower costs.

Based on the foregoing, we believe that the proposed method represents a suitable method for the coordination of the platoons with a virtual hub in a PI based logistics network. However, further research is needed to better exploit the effectiveness of the presented methodology. In the future, it would be advisable to explore the possibility of combining methodologies and the improvements that can be achieved with a larger and more complex network.

5 Conclusion

In the logistics industry, there is a growing trend to reduce carbon emissions, for which platoon is undoubtedly a promising option for the future of freight transport. In this paper, we propose a model that can be used for cooperation of platoons. The model can be adopted in the Physical Internet concept, where communication between platoons can be implemented in an open, global logistics network. The concept enables examination of the vehicle cost changes when platoons reconfigure themselves at a virtual hub by changing vehicles between themselves. Cost includes fuel cost, waiting cost and labour cost.

In the example we compared the basic (non-cooperative) model with the heuristic-based and the RL-based model. We see that the RL-based algorithm is robust, easy to learn and can also deal with stochastic processes using a small network architecture even with a large input state space. We find that with frequent platoon launches, the collaboration is worthwhile, and the reinforcement learning-based model performs better when a large number of vehicles enter the system. One of the future research directions is to further investigate the DQN algorithm performance over a longer period of time and trying other RL algorithms as well. In the future, we also want to extend the model to examine the interaction of multiple nodes on a larger and more complex network. Furthermore, we would also consider optimizing the time interval that determines the platoons launch frequency.

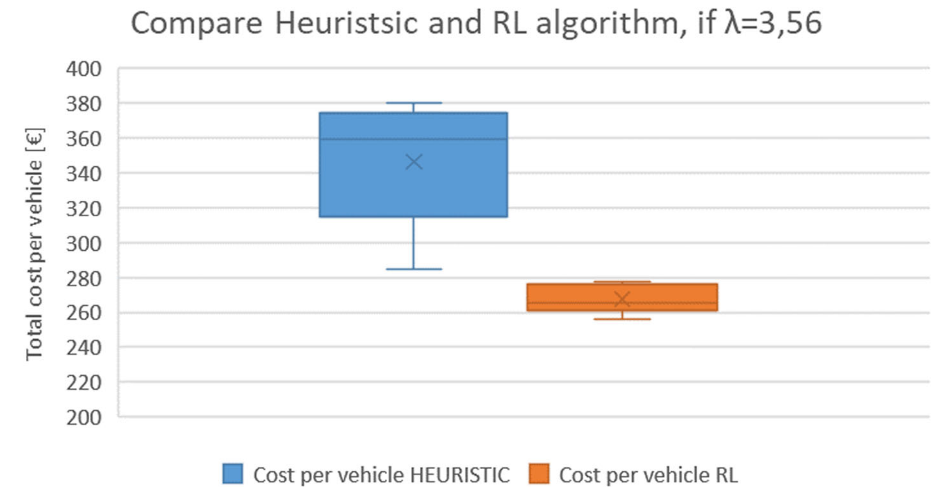

Fig. 6 Compare heuristic algorithm and RL algorithm, if λ is 3,56 min

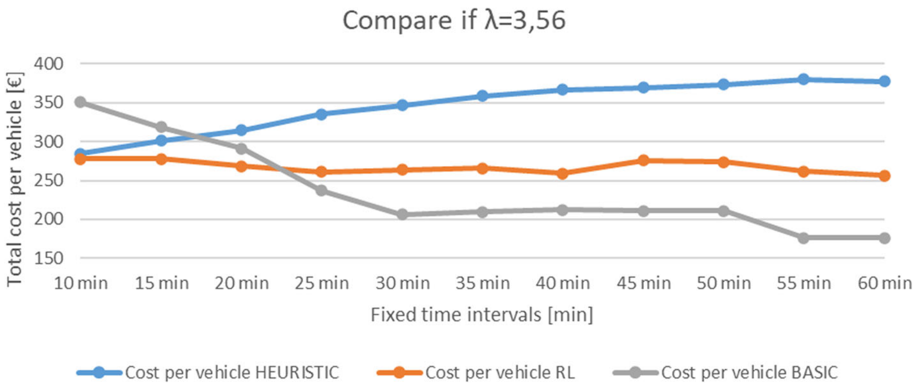

Fig. 7 Compare the 3 models if λ is 3,56 min

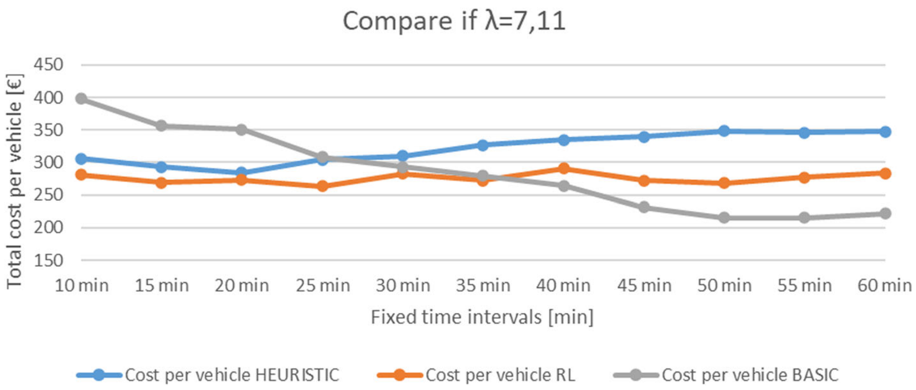

Fig. 8 Compare the 3 models if λ is 7,11 min

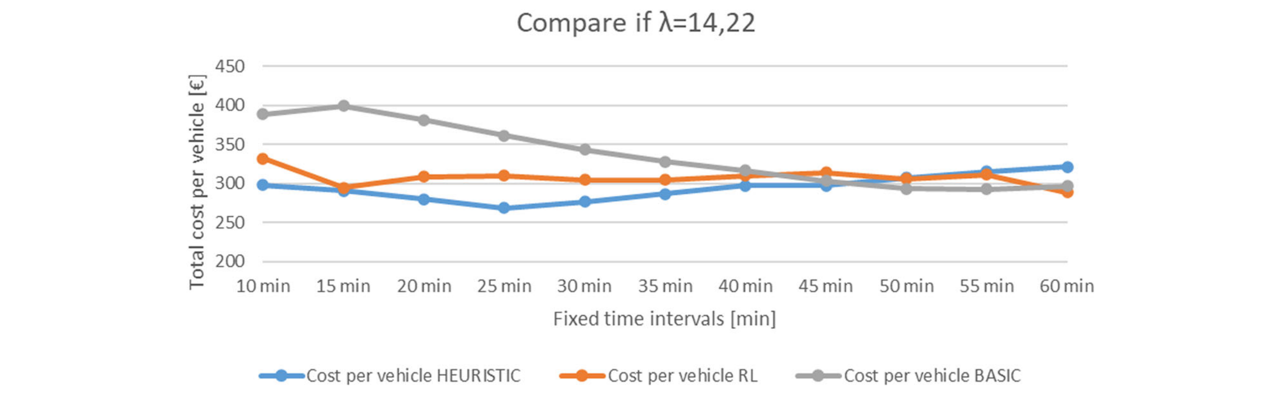

Fig. 9 Compare the 3 models if λ is 14,22 min

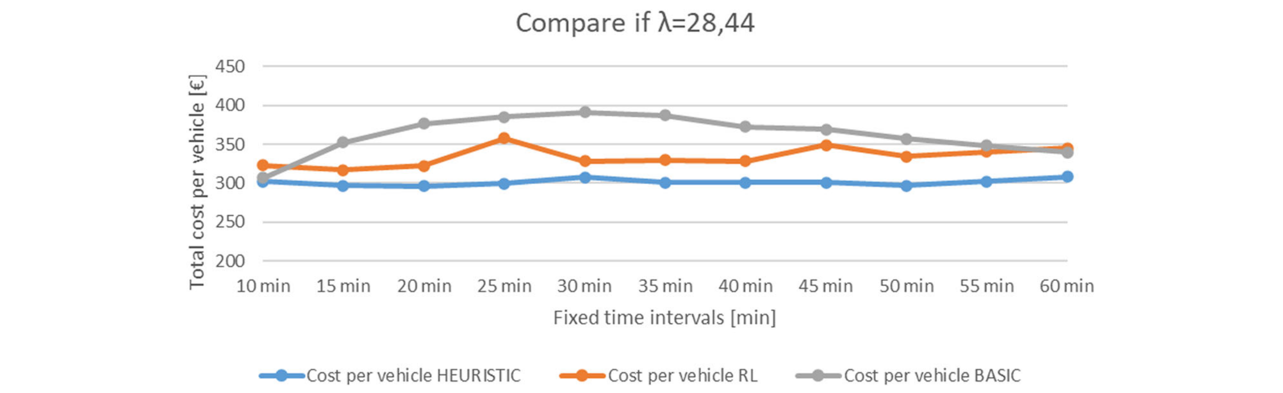

Fig. 10 Compare the 3 models if λ is 28,44 min

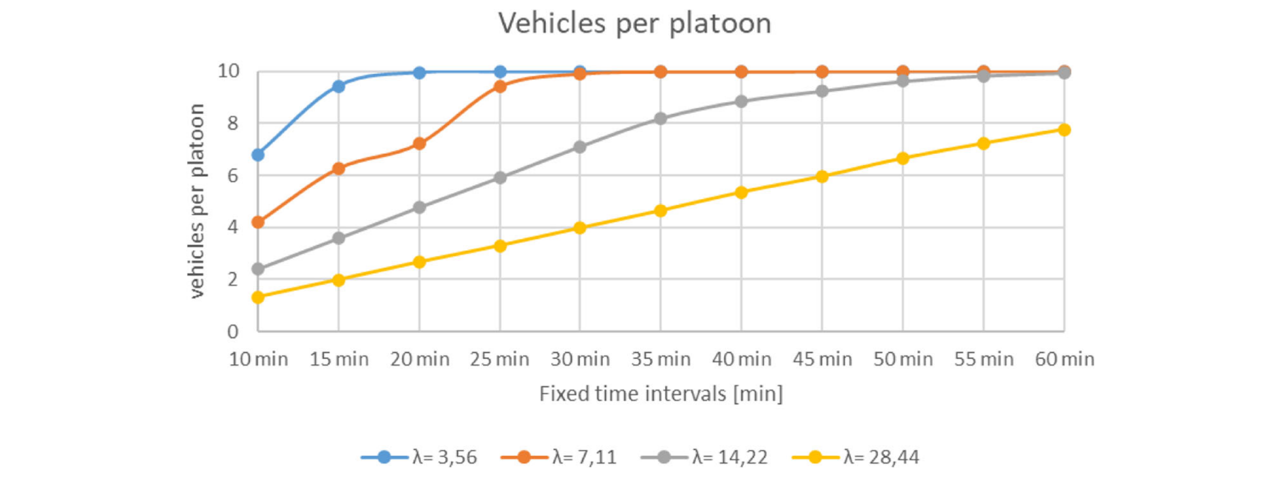

Fig. 11 Compare the number of vehicles per platoon

Acknowledgements
Not applicable.

Authors' contributions
The work was a joint project of all three authors. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The data sets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Budapest University of Technology and Economics, Faculty of Transportation Engineering and Vehicle Engineering, Dept. of Material Handling and Logistics Systems, 1111 Bertalan L. u. 7-9., Building L., Budapest, Hungary. [2]Budapest University of Technology and Economics, Faculty of Electrical Engineering and Informatics, Department of Automation and Applied Informatics, 1117 Magyar tudósok krt. 2. Q.B234, Budapest, Hungary.

Received: 3 December 2019 Accepted: 17 July 2020
Published online: 24 July 2020

References

1. Montreuil, B. (2011). Toward a physical internet: meeting the global logistics sustainability grand challenge. *Logistics Research*, *3*, 71–87.
2. Ciprés, C., & de la Cruz, M.T. (2018). The physical internet from shippers perspective. In B. Müller & G. Meyer (Eds.), *Towards User-Centric Transport in Europe* (pp. 203–221). Switzerland: Springer.
3. Broward, F., McArthur, J., Radovich, C. (2004). Fuel saving achieved in the field test of two tandem trucks. UC Berkeley: California Partners for Advanced Transportation Technology.
4. Maiti, S., Winter, S., Kulik, L. (2017). A conceptualization of vehicle platoons and platoon operations. *Transportation Research Part C: Emerging Technologies*, *80*, 1–19.
5. Schladover, S.E., Nowakowski, C., Lu, X.-Y., Felis, R. (2015). Cooperative adaptive cruise control (cacc) definitions and operating concepts, In *94th TRB Annual Conference January 2015; Washington D. C.* Transportation Research Board (pp. 1–14).
6. Alam, A., Besselink, B., Turri, V., Martensson, J., Johansson, K.H. (2015). Hdv platooning for sustainable freight transportation a cooperative method to enhance safety and efficiency. *IEEE control systems*, *19*, 102–112.
7. Commission, E. Road Fatality Statistics in the EU (info-graphic). https://www.europarl.europa.eu/news/en/headlines/society/20190410STO36615/road-fatality-statistics-in-the-eu-infographic. Accessed 05 Nov 2019.
8. Mihaly, A., & Gaspar, P. (2012). Control of platoons containing diverse vehicles with the consideration of delays and disturbances. *Periodica Polytechnica Transportation Engineering*, *40*(1), 21–26.
9. Janssen, G.R., Zwijnenberg, J., Blankers, I.J., Kruijff, J.S. (2015). Truck platooning: Driving the future of transportation. *TNO Whitepaper*, *36*, 1–36.
10. Qiao, B., Pan, S., Ballot, E. (2018). Revenue optimization for less-than-truckload carriers in the physical internet: Dynamic pricing and request selection. *Computers & Industrial Engineering*, *30*(7), 2631–643.
11. Commission, E. Horizon 2020 - Work Programme 2018-2020 Smart, Green and Integrated Transport. https://ec.europa.eu/programmes/horizon2020/en/h2020-section/smart-green-and-integrated-transport. Accessed 12 Nov 2019.
12. Delgado, O., Rodríguez, F., Muncrief, R. (2017). Fuel efficiency technology in european heavy-duty vehicles: Baseline and potential for the 2020–2030 time frame. International Council on Clean Transportation.
13. Commission, E. Reducing CO2 Emissions from Heavy-duty Vehicles. https://ec.europa.eu/clima/policies/transport/vehicles/heavy_en. Accessed 12 Nov 2019.
14. Larson, J., Liang, K.Y., Johansson, K.H. (2015). A distributed framework for coordinated heavy-duty vehicle platooning. *IEEE Transactions on Intelligent Transportation Systems*, *16*, 419–429.
15. Alkim, T., Vliet, A., Aarts, L., Eckhardt, J. (2016). European truck platooning challenge 2016 - creating next generation mobility: Lessons learnt. *Technical Report*. Ministry of Infrastructure and the Environment, Netherlands.
16. Bhoopalam, A.K., Agatz, N., Zuidwijk, R. (2018). Planning of truck platoons: A literature review and directions for future research. *Transportation Research Part B: Methodological*, *107*, 212–228.
17. van de Hoef, S., Johansson, K.H., Dimarogonas, D.V. (2017). Fuel-efficient en route formation of truck platoons. *IEEE Transactions on Intelligent Transportation Systems*, 34–56. https://doi.org/10.1109/tits.2017.2700021.
18. Zhang, W., Jenelius, E., Ma, X. (2017). Freight transport platoon coordination and departure time scheduling under travel time uncertainty. *Transportation Research Part E: Logistics and Transportation Review*, *98*, 1–23.
19. Adler, A., Miculescu, D., Karaman, S. (2016). Optimal policies for platooning and ride sharing in autonomy- enabled transportation. *Workshop on Algorithmic Foundations of Robotics (WAFR)*, 848–63.
20. Meisen, P., Seidl, T., Henning, K. (Eds.) (2008). *A Data-Mining Technique for the Planning and Organization of Truck Platoons 2008*. Paris: International Conference on Heavy Vehicles.
21. Boysen, N., Briskorn, D., Schwerdfeger, S. (2018). The identical-path truck platooning problem. *Transportation Research Part B Methodological*, *109*, 26–39.
22. Larsen, R., Rich, J., Rasmussen, T.K. (2019). Hub-based truck platooning: Potentials and profitability. *Transportation Research Part E: Logistics and Transportation Review*, 1–25. https://doi.org/10.1016/j.tre.2019.05.005.
23. Alshiddi, R.S., & Hexmoor, H. (Eds.) (2019). *Path Finding and Joining for Truck Platoons 2019, USA*. USA: 2019 International Conference on Artificial Intelligence.
24. Puskás, E., & Bohács, G. (2018). Concepting freight holding problems for platoons in physical internet systems. *Acta logistica -International Scientific Journal about Logistics*, *6*, 19–27.
25. Meyer, G., & Beiker, S. (2015). *Road Vehicle Automation 2*. Switzerland: Springer.
26. Muncrief, R. (2017). Shell game? debating real-world fuel consumption trends for heavy-duty vehicles in Europe. *International Council on Clean Transportation*, *5*, 1–5.
27. Davila, A. (2013). *SARTRE Report on fuel consumption. Technical Report for European Commission under the Framework 7 Programme Project 233683 Deliverable 4.3*. Cambridge: Ricardo UK Limited.
28. Lammert, M.P., Duran, A., Diez, J., Burton, K., Nicholson, A. (Eds.) (2014). *Effect of Platooning on Fuel Consumption of Class 8 Vehicles Over a Range of Speeds, Following Distances, and Mass 2014 Illinois*. Illinois: SAE 2014 Commercial Vehicle Engineering Congress (COMVEC).
29. Sokolov, V., Larson, J., Munson, T., Auld, J., Karbowski, D. (2017). Platoon formation maximization through centralized routing and departure time coordination. *Transportation Research Record Journal of the Transportation Research Board*, 1–14. https://doi.org/10.3141/2667-02.
30. Zhang, W., Sundberg, M., Karlström, A. (2017). Platoon coordination with time windows: an operational perspective. *Transportation Research Procedia*, *27*, 357–364.
31. Commission, E. Analysis of the Contribution of Transport Policies to the Competitiveness of the EU Economy and Comparison with the United States. https://ec.europa.eu/ten/transport/studies/doc/compete/compete_report_en.pdf?fbclid=IwAR1AnE5VWBxOVBcsAt3y4Fb0aRGM62_zX9tptz4m_o5L7iRKl-7uAFuFE3A. Accessed 04 Nov 2019.
32. des Transports Routiers, U.I. Selected Recent Statistics on Road Freight Transport in Europe. https://www.iru.org/sites/default/files/2016-01/en-statistics-goods_0.pdf?fbclid=IwAR0ZTl7rM97s6s7of__sietJ5YWs4kosujfVt7QPgsllEqOBCSLgPfzgEE8. Accessed 05 Nov 2019.
33. Pontrandolfo, P., Gosavi, A., Gosavi, A., Okogbaa, O.G., Das, T.K. (2002). Global supply chain management: A reinforcement learning approach. *International Journal of Production Research*, *40*(6), 1299–317.
34. Min, H. (2009). Artificial intelligence in supply chain management: Theory and applications. *International Journal of Logistics Research and Applications*, *13*, 13–39.
35. Stockheim, T., & Schwind, M. (2002). WoKoenig: A reinforcement learning approach for supply chain management. *International Journal of Production Research*, *40*, 1–14.
36. Habib, A., Khan, M.I., Uddin, J. (2016). Optimal route selection in complex multi-stage supply chain networks using sarsa (λ). *9th International Conference on Computer and Information Technology (ICCIT)*, 170–75.

37. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *CoRR*, *abs/1312.5602*.
38. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature Journal*, *529*, 484–89.
39. Sutton, R. (1992). The challenge of reinforcement learning. *Machine Learning*, *8*, 225–227.
40. Sutton, R., & Barto, A. (2018). *Reinforcement learning: An introduction*. Massachusetts: MIT Press.
41. Mnih, V., et al. (2015). Human-level control through deep reinforcement learning. *Nature Journal*, *518*, 529–33.
42. Wang, Z., Schaul, T., Hessel, M., Lanctot, M., de Freitas, N. (2015). Dueling network architectures for deep reinforcement learning. *CoRR*, *abs/1511.06581*(48), 1995–2003.
43. Williams, R.J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*, 229–256.
44. Williams, R.J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*, 229–56.
45. Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W. (2016). Openai gym, *CoRR*. arXiv:1606.01540.
46. Puskás, E., Budai, A., Bohacs, G. Reinforcement Learning Based Model for Platoons. https://github.com/adamtiger. Accessed 05 Nov 2019.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.