Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 333 (2004) 379–440
PHYSICA A
www.elsevier.com/locate/physa

Three-phase traffic theory and highway capacity

Boris S. Kerner

Daimler Chrysler AG, RIC/TS, T729, 70546 Stuttgart, Germany

Received 28 July 2003

Abstract

Hypotheses and some results of the three-phase traffic theory by the author are compared with results of the fundamental diagram approach to traffic flow theory. A critical discussion of model results about congested pattern features which have been derived within the fundamental diagram approach to traffic flow theory and modeling is made. The empirical basis of the three-phase traffic theory is discussed and some new spatial–temporal features of the traffic phase "synchronized flow" are considered. A probabilistic theory of highway capacity is presented which is based on the three-phase traffic theory. In the frame of this theory, the probabilistic nature of highway capacity in free flow is linked to an occurrence of the first order local phase transition from the traffic phase "free flow" to the traffic phase "synchronized flow". A numerical study of congested pattern highway capacity based on simulations of a KKW cellular automata model within the three-phase traffic theory is presented. A congested pattern highway capacity which depends on features of congested spatial–temporal patterns upstream of a bottleneck is studied.

1. Introduction

Real traffic is a dynamical process which occurs both in space and time. This spatial–temporal process shows very complex dynamical behavior. In particular, highway traffic can be either free or congested. Congested traffic states can be defined as the traffic states where the average vehicle speed is lower than the minimum possible average speed in free flow (e.g., Ref. [1]). It is well known that in contrast to free traffic flow, in congested traffic a collective behavior of vehicles plays an important role (the collective flow by Prigogine and Herman [2]), and a synchronization of vehicle speeds across different highway lanes usually occurs [1]. In congested traffic, complex spatial–temporal patterns are observed, in particular a sequence of moving traffic jams, the so-called "stop-and-go" phenomenon (e.g., the classical works by Treiterer [3] and Koshi et al. [1]). Recall that *a moving jam* is a moving localized structure. The moving jam is spatially restricted by two upstream moving jam fronts where the vehicle speed

0378-4371/$ - see front matter
doi:10.1016/j.physa.2003.10.017

and the density change sharply. The vehicle speed is low (sometimes as low as zero) and the density is high inside the moving jam.

Congested traffic usually occurs at a highway bottleneck, e.g., at the bottleneck due to an on-ramp. In empirical investigations, the onset of congested traffic is accompanied by the breakdown phenomenon, i.e., by a sharp decrease in the vehicle speed at the bottleneck (see e.g., papers by Athol and Bullen [4], Banks [5,6], Hall et al. [7,8], Elefteriadou et al. [9], Kerner and Rehborn [10] and by Persaud et al. [11]). It has been found that the breakdown phenomenon has a probabilistic nature, i.e., the probability of the speed breakdown is an increasing function of the flow rate in free flow at the bottleneck [11]. Besides, it has been found that the capacity of congested bottleneck, i.e., highway capacity after the breakdown phenomenon at the bottleneck has occurred is often lower than the capacity in free flow before—the so-called phenomenon "capacity drop" [5–8].

Concerning the important role of highway bottlenecks it should be noted that although congested traffic can occur away from bottlenecks [12], they have an important impact just like defects in physical systems which can play an important role for the phase transitions and for the formation of spatial–temporal patterns. The role of the bottlenecks in traffic flow is as follows: Congested traffic occurs most frequently at highway bottlenecks (e.g., Refs. [13,14]). The bottlenecks can result from for example due to road works, on and off ramps, a decrease in the number of highway lanes, road curves and road gradients.

Although the complexity of traffic is linked to the occurrence of spatial–temporal patterns, some of the traffic features can be understood if *average* traffic characteristics are considered. Thus, important empirical methods in traffic science are empirical flow–density and speed–density relationships which are related to measurements of some *average* traffic variables at a highway location, in particular at a highway bottleneck. The empirical relationship of the *average* vehicle speed on the vehicle density must be related to an obvious result observed in real traffic flow: the higher the vehicle density, the lower the average vehicle speed. When the flow rate, which is the product of the vehicle density and the average vehicle speed, is plotted as a function of the vehicle density one gets what is known as the empirical fundamental diagram. It must be noted that the empirical fundamental diagram is successfully used for different important applications where some average traffic flow characteristics should be determined (e.g., Refs. [14,15]).

Recently dependencies of the empirical fundamental diagram on the type of the congested pattern at a freeway bottleneck and on the freeway location where the empirical fundamental diagram is measured have been found [16].

Over the past 80 years scientists have developed a wide range of different mathematical models of traffic flow to understand these complex non-linear traffic phenomena (see the classical papers by Lighthill and Whitham [17], Richards [18], Herman et al. [19], Gazis et al. [20], Kometani and Sasaki [21], Newell [22], Prigogine [23], Payne [24], Gipps [25], the books by Leuzbach [26], May [14], Daganzo [13], Prigogine and Herman [2], Wiedemann [27], Whitham [28], Cremer [29], Newell [30], the reviews by Chowdhury et al. [31], Helbing [32], Nagatani [33], Nagel et al. [34] and the conference proceedings [35–41]). Clearly these models must be based on the real

behavior of drivers in traffic, and their solutions should show phenomena observed in real traffic.

1.1. Fundamental diagram approach to traffic flow theory and modeling

Up to now by a development of a mathematical traffic flow model which should explain empirical spatial–temporal congested patterns, it has been self-evident that hypothetical *steady*-state solutions of the model should belong to a curve in the flow–density plane (see, e.g., [17,20,22,23,2,38–40,42–57,28,29,26,58–70] and the recent reviews [31–34]). The above term *steady state* designates the hypothetical model solution where vehicles move at the same distances to one another with the same time-independent vehicle speed. Therefore, steady states are hypothetical spatially homogeneous and time-independent traffic states (steady states are also often called "homogeneous" or "equilibrium" model solutions; we will use in the article for these hypothetical model traffic states the term "steady" states or "steady speed" states). The curve in the flow–density plane for steady-state model solutions goes through the origin and has at least one maximum. This curve is called *the fundamental diagram* for traffic flow.

The postulate about the fundamental diagram underlies almost all traffic flow modeling approaches up to now (see reviews [31–34]) in the sense that the models are constructed such that in the *unperturbed*, *noiseless limit* they have a fundamental diagram of steady states, i.e., the steady states form a curve in the flow–density plane. The fundamental diagram is either a result of the model (e.g., for the models by Gazis et al. [20], Gipps [25] Nagel–Schreckenberg cellular automata (CA for short) [42–44,52,69], Krauß et al. [51], Helbing, Treiber and co-workers [56,66,71], Tomer, Halvin and co-worker [68,72]) or the fundamental diagram is hypothesized in a model (e.g., the optimal velocity models by Newell [22], Whitham [45], Bando, Sugiyama and co-workers [47,48] and the macroscopic model by Payne [24]). Moving jams which are calculated from these models for a homogeneous road (i.e., a road without bottlenecks) are due to the instability of steady states of the fundamental diagram within some range of vehicle densities (see reviews [31–34]). This is one of the reasons why we find it helpful to classify these models as belonging to what we call the "fundamental diagram approach".

In 1955 Lighthill and Whitham [17] wrote in their classical work (see p. 319 in [17]): "... The fundamental hypothesis of the theory is that at any point of the road the flow (vehicles per hour) is a function of the concentration (vehicles per mile)...". Apparently the empirical fundamental diagram was the reason that the fundamental diagram approach has already been introduced in the first traffic flow models derived by Lighthill and Whitham [17], by Gasis et al. [20], and by Newell [22].

Concerning the theoretical fundamental diagram, it must be noted that in real congested traffic complex spatial–temporal traffic patterns are observed (e.g., Refs. [1,3]). These patterns are spatially non-homogeneous. This spatial behavior of congested patterns is a complex function of time. An averaging of traffic variables related to congested patterns over long enough time intervals gives a relation between different *averaged* vehicle speeds and densities. Thus, the empirical fundamental diagram is related to *averaged* characteristics of spatial–temporal congested patterns measured at a

highway location rather than to features of the hypothetical steady states of congested traffic on the theoretical fundamental diagram. For this reason, the existence of the theoretical fundamental diagram is only *a hypothesis*. This is also confirmed by a recent empirical study where it has been found that the empirical fundamental diagram strongly depends on the congested pattern type and on the freeway location [16].

It has recently been found that empirical features of the phase transitions in traffic flow and most of empirical spatial–temporal pattern features [74–76] are qualitatively different from those which follow from mathematical traffic flow models in the fundamental diagram approach which is considered in the reviews [31–34].

1.2. Three-phase traffic theory

For this reason in 1996–2000 the author, based on an empirical traffic flow analysis, introduced a concept called "synchronized flow" and the related three-phase traffic theory [12,74,75,77–84].

1.2.1. The concept "synchronized flow"

In the concept "synchronized flow", there are two qualitatively different phases, the traffic phase called "synchronized flow" and the traffic phase called "wide moving jam", which should be distinguished in congested traffic [74,75,77–80]. This distinguishing is based on qualitatively different *empirical spatial–temporal* features of these phases. Traffic consists of free flow and congested traffic. Congested traffic consists of two traffic phases. Thus, there are three traffic phases:

(1) Free flow.
(2) Synchronized flow.
(3) Wide moving jam.

In the three-phase traffic theory, features of spatial–temporal congested patterns are explained based on the phase transitions between these three traffic phases.

Objective criteria to distinguish between the traffic phase "synchronized flow" and the traffic phase "wide moving jam" are based on qualitative different *empirical spatial–temporal* features of these two traffic phases. A wide moving jam is a moving jam which possesses the following characteristic feature. Let us consider the downstream front of the wide moving jam where vehicles accelerate escaping from a standstill inside the wide moving jam. This downstream jam front propagates on a highway *keeping* the mean velocity of this front. As long as a moving jam is a wide moving jam this characteristic effect—the keeping of the velocity of the downstream jam front—remains even if the wide moving jam propagates through any complex traffic states and through any highway bottlenecks. In contrast, the downstream front of the traffic phase "synchronized flow" (where vehicles accelerate escaping from synchronized flow to free flow) is usually fixed at the bottleneck. Corresponding to the definition of the traffic phase "wide moving jam" and to the concept "synchronized flow" where there are only two traffic phases in congested traffic, any state of congested traffic which does not possess the above characteristic feature of a wide moving jam is related to the traffic state "synchronized flow". A more detailed consideration of the objective criteria of traffic phases and empirical examples of the application of these objective criteria for

the determination of the phase "synchronized flow" and the phase "wide moving jam" in congested traffic can be found in Ref. [76]. It must be noted that there are at least two well-known empirical effects in congested traffic (e.g., [1]): (1) Synchronization of the average vehicle speed between different freeway lanes. (2) A wide spreading of empirical data in the flow-density plane. These effects can occur in both traffic phases of congested traffic, "synchronized flow" and "wide moving jam". The distinguishing between these traffic phases is made only based on the above objective criteria for traffic phases rather than on the speed synchronization effect or on the effect of the wide spreading of empirical data in the flow-density plane.

It should be stressed that the concept "synchronized flow" and the related methodology of the congested pattern study, which has been used for the definition of the three traffic phases below, is based on an analysis of *empirical spatial–temporal* features of congested patterns [74,77,78] rather than on a dynamical analysis of data (e.g., in the flow–density plane) only. First, a spatial–temporal study of traffic must be made. Only after the traffic phases "synchronized flow" and "wide moving jam" have already been distinguished, based on the objective criteria for the traffic phases discussed above some of the pattern features can further be studied, e.g., in the flow–density plane. In particular, this procedure has already been used in Ref. [77]: At the first step, a spatial–temporal analysis of empirical data has been made and the phases "synchronized flow" and "wide moving jam" have been identified. At the next step, the traffic phase "synchronized flow" has been plotted in the flow–density plane without any wide moving jams and some of the features of "synchronized flow" have been studied. Measured data on sections of the highway A5 in Germany which have been used in Refs. [12,74–84] also comprise the information about vehicle types (number of vehicles and long vehicles) and individual vehicle speeds passing the detector during each one minute interval of averaging. Using the latter information in addition to the spatial–temporal data analysis of 1 min averaged data, the determination of the type of synchronized flow in empirical studies of synchronized flow (the type (i), or (ii) or else (iii) [77]) have been made. Besides, the individual vehicle speeds allow us to answer the question of whether there are narrow moving jams in synchronized flow or not. All these steps of data analysis have been made in Refs. [12,74–84]. We will illustrate this in Section 2 where an empirical example of a synchronized flow pattern will be studied.

1.2.2. The fundamental hypothesis of the three-phase traffic theory

Another claim of the concept of "synchronized flow" is the hypothesis about steady states of synchronized flow. This is the fundamental hypothesis of the three-phase traffic theory. The fundamental hypothesis of the three-phase traffic theory reads as follows [74,78–80]:

Hypothetical steady states of synchronized flow cover a two-dimensional region in the flow–density plane (Fig. 1). This means that in these hypothetical steady states of synchronized flow, where all vehicles move at the same distance to one another and with the same time-independent speed, a given steady vehicle speed is related to an infinite multitude of different vehicle densities and a given vehicle density is related to an infinite multitude of different steady vehicle speeds. This hypothesis means that there

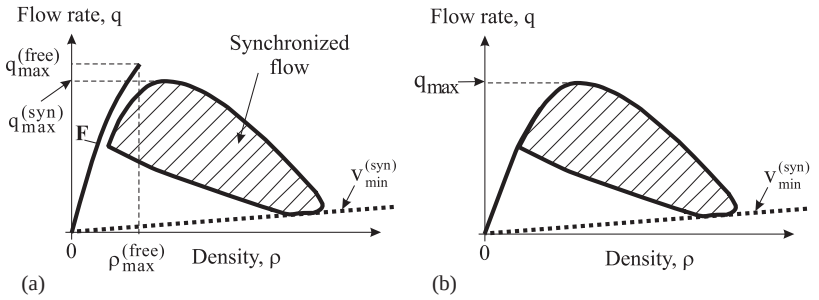

Fig. 1. The fundamental hypotheses of the three-phase traffic, theory [74,78]: States of free flow (curve F) and spatially homogeneous and time-independent states (steady states) of synchronized flow (dashed region) in the flow density plane for a multi-lane (in one direction) homogeneous (without bottlenecks) road in the three-phase traffic theory (a). For a comparison, in (b) steady states for a homogeneous (without bottlenecks) one-lane road are shown. The dotted line is related to a minimum possible speed in steady states of synchronized flow $v_{\min}^{(\mathrm{syn})}$ [75,82].

is *no* fundamental diagram for hypothetical steady speed states of synchronized flow. This hypothesis has recently been used in a microscopic three-phase traffic flow theory [85–87]. It occurs that this theory, which is also based on other hypotheses of the three-phase traffic theory (Section 4), explains and predicts main features of empirical phase transitions and spatial–temporal congested patterns found in Refs. [74,76].

The fundamental hypothesis of the three-phase traffic theory is therefore in contradiction with the hypothesis about the existence of the fundamental diagram for hypothetical steady states of mathematical models and theories in the fundamental diagram approach. An explanation of the fundamental hypothesis of the three-phase traffic theory will be done in Section 4.1.1.

1.2.3. Explanation of the terms "synchronized flow" and "wide moving jam"

The term "synchronized flow" should reflect the following features of this traffic phase: (i) It is a non-interrupted traffic flow rather than a long enough standstill as it usually occurs inside a wide moving jam. The word "flow" should reflect this feature. (ii) There is a *tendency* to a synchronization of vehicle speeds across different lanes on a multi-lane road in this flow. Besides, there is a *tendency* to a synchronization of vehicle speeds on each of the road lanes (a bunching of the vehicles) in synchronized flow due to a relatively low mean probability of passing in synchronized flow. The word "synchronized" should reflect these speed synchronization effects.

The term "wide moving jam" should reflect the characteristic feature of the jam to propagate through any other states of traffic flow and through any bottlenecks keeping the velocity of the downstream jam front. The word combination "moving jam" should reflect the feature of the jam *propagation* as a whole localized structure on a road. If the width of a moving jam is considerably higher than the widths of the jam fronts and the speed inside the jam is zero then the moving jam possesses this characteristic feature. The word "wide" (the jam width in the longitudinal direction) should reflect this characteristic feature of the jam propagation *keeping* the velocity of the downstream jam front. However, the distinguishing between the traffic phases "synchronized flow"

and "wide moving jam" is made only based on the objective criteria for traffic phases which have been considered in Section 1.2.1.

This article is organized as follows. In Section 2, firstly an overview of known features of synchronized flow is made (Section 2.1) and then new empirical results about spatial–temporal features of synchronized flow are presented. These empirical results should help to understand some of the hypotheses to the three-phase traffic theory. A critical analysis of the application of the fundamental diagram approach for a description of phase transitions and of spatial–temporal features of congested patterns will be made in Section 3. In Section 4, a comparison of already known hypotheses to the author's three-phase traffic theory with results of the fundamental diagram approach to traffic flow theory will be considered. In this section, we will also consider some new hypotheses about a Z- and *double* Z-shaped characteristics of traffic flow. The double Z-characteristic should explain phase transitions which are responsible for the wide moving jam emergence in real traffic flow. Other new results are presented in Section 5 where a probabilistic theory of highway capacity which is based on the three-phase traffic theory is considered. This general theory will be illustrated and confirmed by new numerical results of a study of a KKW cellular automata model within the three-phase traffic theory.

2. Empirical features of synchronized flow

In this section some new empirical features of synchronized flow will be considered. These features will be used below for an explanation of the three-phase traffic theory. In particular, these results allow us to give the empirical basis for the hypothesis about a Z-shape of the probability of passing in traffic flow as a function of the density. However, firstly a brief overview of empirical features of synchronized flow is made.

2.1. Main empirical features of phase transitions, onset of congestion and congested patterns at bottlenecks (overview)

Moving jams do not emerge in free flow, if synchronized flow is not hindered [75]. Instead, the moving jams emerge due to a sequence of two first-order phase transitions [74]: First the transition from free flow to synchronized flow occurs (it is called the F $\rightarrow$ S transition) and only later and usually at a different highway location wide moving jams emerge in the synchronized flow (the latter transition is called the S $\rightarrow$ J transition and the sequence of both transitions is called the F $\rightarrow$ S $\rightarrow$ J transitions).

In particular, the onset of congestion at a bottleneck, i.e., the well-known breakdown phenomenon in free flow is linked to the F $\rightarrow$ S transition in an initial free flow rather than to the wide moving jam emergence (the F $\rightarrow$ J transition) [74,76]. This means that at the same density in free flow at the bottleneck the probability of the F $\rightarrow$ S transition should be considerably higher than the probability of the F $\rightarrow$ J transition.

Empirical investigations show [76,103] that there are two main types of congested patterns at an isolated bottleneck: [1]

[1] Note that an isolated bottleneck is a highway bottleneck which is far enough away from other bottlenecks where congested patterns can emerge.

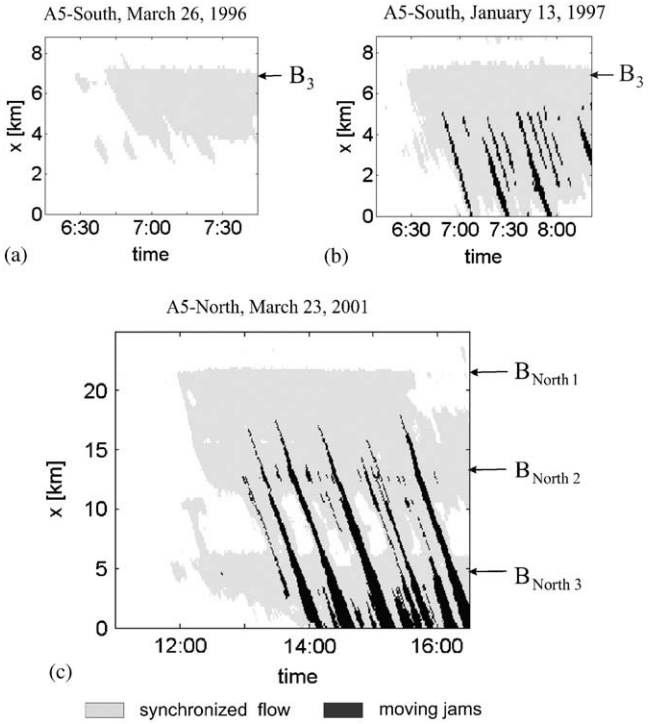

Fig. 2. Empirical congested pattern classification [76]. Examples of empirical congested patterns in space and time: (a) Synchronized flow pattern (SP). (b) General pattern (GP). (c) Expanded congested pattern (EP). Infrastructure of the section of the freeway A5-South in Germany is shown in Fig. 3 in Ref. [76]. The scheme of the section on the freeway A5-North is shown in Fig. 3(a). Highway bottlenecks B_3 on A5-South and $B_{\text{North }1}$, $B_{\text{North }2}$, $B_{\text{North }3}$ on A5-North have been considered in Ref. [76]. "White"—free flow, "gray"—synchronized flow, "black"—moving jams.

The synchronized flow pattern or *SP* for short (Fig. 2(a)): The SP consists of synchronized flow upstream of the isolated bottleneck *only*, i.e., *no* wide moving jams emerge in that synchronized flow.

The general pattern or *GP* for short (Fig. 2(b)): The GP is the congested pattern at the isolated bottleneck where synchronized flow occurs upstream of the bottleneck and wide moving jams spontaneously emerge in that synchronized flow. Thus the GP consists of both traffic phases in congested traffic: "synchronized flow" and "wide moving jam".

However, dependent on the bottleneck features and on traffic demand, the GP and the SP show a diverse variety of special cases. In particular, there are three main different types of synchronized flow patterns (SP) at the isolated bottleneck: The localized SP (LSP), i.e., the SP whose width is spatially limited over time (Fig. 2(a)),[2] the widening

[2] LSP can also be called a "shortening" SP. In the article, following the empirical congested pattern study [76] we will use the term LSP.

SP (WSP), i.e., the SP whose width is continuously widening over time (see empirical example below) and the moving SP (MSP), i.e., the SP which propagates as a whole localized pattern on the road (see empirical example in Fig. 6(a) in Ref. [76]).

The usual scenario of the GP emergence is the following [74,76]: First, the $F \to S$ transition occurs at a highway bottleneck. The downstream front of the synchronized flow (the front where drivers escape from the synchronized flow upstream to the free flow downstream) is fixed at the bottleneck. The upstream front of the synchronized flow (this front separates free flow upstream from the synchronized flow downstream) propagates upstream. Later in this synchronized flow upstream of the bottleneck narrow moving jams spontaneously emerge. Narrow moving jams propagate upstream and self-grow.[3] Finally, the growing narrow moving jams (or only a part of them) transform into wide moving jams. This transformation is related to the $S \to J$ transition.

It must be noted that the well-known and very old term "stop-and-go" traffic which is related to a sequence of moving traffic jams will not be used in this article. This is linked to the following: For a traffic observer, both a sequence of narrow moving jams and a sequence of wide moving jams is "stop-and-go" traffic. However, as it has already been mentioned (see footnote 3) narrow moving jams belong to the traffic phase "synchronized flow" whereas wide moving jams belong to the qualitatively different traffic phase "wide moving jam".

It has been found that moving jams are most to emerge in dense synchronized flow of lower average vehicle speed [74,76]. In other words, the average speed in synchronized flow of a GP where a wide moving jam emerges is usually considerably lower than the speed in synchronized flow of a SP. The dense synchronized flow in GP appears due to the pinch effect in the synchronized flow, i.e., due to a strong compression of the synchronized flow upstream of the bottleneck. Besides, the lower the average speed in the pinch region of synchronized flow in a GP the higher is the frequency of the moving jam emergence in this synchronized flow. In particular, the average speed in the pinch region of synchronized flow is decreasing upstream of an on-ramp if the flow rate to the on-ramp is increasing. Consequently the frequency of the moving jam emergence increases when the flow rate to the on-ramp increases. This is because the average speed in synchronized flow upstream of the on-ramp decreases. Even at the highest observed flow rate to the on-ramp moving jams emerge (with the highest frequency) in synchronized flow upstream of the on-ramp.

In contrast, empirical investigations of synchronized flow allow us to suggest that in synchronized flow of higher vehicle speed moving jams do not necessarily emerge [74,76]. In this case, a SP can occur where no wide moving jams emerge [76,85]. One of the SPs is a WSP. The WSP should occur when a bottleneck introduces a relatively small disturbance for traffic flow.

[3] A narrow moving jam is a moving jam which does not possess the characteristic feature of a wide moving jam to keep the downstream jam front velocity propagating through any state of traffic and through any bottlenecks. Narrow moving jams belong to congested traffic. Corresponding to the objective criteria of traffic phases in congested traffic any state of congested traffic which does not possess this characteristic feature of a wide moving jam is related to the traffic phase "synchronized flow". Thus, narrow moving jams are states of the traffic phase "synchronized flow" [76].

This case which can be observed at a bottleneck due to off-ramps will be considered below. It will be shown, that in this case synchronized flow in a WSP can indeed exist upstream of the bottleneck on a long stretch of the highway (about 4.5 km) during a long time (more than 60 min) without wide moving jam emergence in that synchronized flow. This confirms the above suggestion made that in synchronized flow of higher vehicle speed moving jams do not necessarily emerge.

The GP and the SP which have briefly been discussed above appear at an "isolated" bottleneck. An influence of the other possible freeway bottlenecks on the pattern formation at the isolated bottleneck should be negligible.

On real highways there are a lot of bottlenecks where different congested patterns almost simultaneously can emerge. If two such bottlenecks exist close to one another, then an expanded congested pattern (EP) can be formed. In the EP, synchronized flow covers at least two bottlenecks. For example, let us consider two bottlenecks which are close to one another. The bottleneck downstream will be called "downstream bottleneck" and the bottleneck upstream will be called "upstream bottleneck". Let us assume that the $F \to S$ transition occurs at the downstream bottleneck, i.e., a synchronized flow emerges at this bottleneck. Due to the upstream propagation of the synchronized flow, this flow can reach the upstream bottleneck. In this case, the synchronized flow can propagate upstream of the upstream bottleneck covering both bottlenecks: an EP appears.

Another empirical example which will be considered below is the following: If a WSP occurs at the downstream bottleneck, then due to the continuous upstream widening of synchronized flow in the WSP the upstream front of the synchronized flow always at some time reaches the upstream bottleneck. After the synchronized flow is upstream of the upstream bottleneck, the synchronized flow covers both bottlenecks, i.e., the initial WSP transforms into an EP. From this example we may conclude that the congested pattern at the downstream bottleneck can be considered as a WSP in only a finite time interval as long as the synchronized flow of the WSP does not reach the upstream bottleneck. After the upstream front of synchronized flow of the WSP has reached the upstream bottleneck, an EP appears. Synchronized flow of this EP consists of the synchronized flow of the initial WSP and the synchronized flow at the upstream bottleneck. The upstream bottleneck may make a great influence on the synchronized flow of the initial WSP, e.g., this bottleneck can lead to the wide moving jam formation in the synchronized flow. This occurs in the example of this WSP which will be considered in Section 2.2.

There can be a lot of different types of EPs. As it has been shown in Ref. [76], EPs can be explained within the three-phase traffic theory. This conclusion is linked to the empirical fact [76] that congested traffic of all known EP types can consist of the traffic phase "wide moving jam" and the traffic phase "synchronized flow" only. This result of the empirical study [76] is illustrated in Fig. 2 where an example of an EP is shown. Firstly, the synchronized flow occurs at the downstream bottleneck. The upstream front of synchronized flow propagates upstream and reaches the first upstream bottleneck and later the second upstream bottleneck: The EP appears where synchronized flow covers these three bottlenecks. The location of the downstream bottleneck is marked $B_{\mathrm{North}\ 1}$ and the locations of two upstream bottlenecks are marked $B_{\mathrm{North}\ 2}$ and $B_{\mathrm{North}\ 3}$ in

Fig. 2(c), respectively. These are effective bottlenecks on the freeway A5-North which have been considered in Ref. [76].

We can see in Fig. 2(c) that congested traffic in the EP consists of the traffic phase "wide moving jam" and the traffic phase "synchronized flow": Some of the moving jams, which have initially emerged upstream of the downstream bottleneck ($B_{\text{North 1}}$), propagate through the first upstream bottleneck ($B_{\text{North 2}}$) keeping the velocity of the downstream jam front. These moving jams belong to the traffic phase "wide moving jam". Later on these wide moving jams propagate also through the second upstream bottleneck on this freeway section ($B_{\text{North 3}}$). During the wide moving jam propagation the width of the jams increases over time (Fig. 2(c)). These empirical results are also confirmed by a microscopic three-phase traffic theory of EPs which has recently been developed in Ref. [87].

2.2. Empirical example of WSP

The empirical study made in Ref. [76] allows us to assume that WSP can occur at off-ramps. The importance of the WSP analysis is linked to the possibility to show empirical features of synchronized flow of a relatively high vehicle speed.

An example of a WSP which occurs upstream of the off-ramp D25-off on the section of the highway A5-North in Germany (the State "Hessen") is shown in Figs. 3–6. The section of the highway A5-North (Fig. 3(a)) has already been described in Ref. [76] (see Fig. 3(c) in this paper). A WSP occurs as a result of the F $\rightarrow$ S transition upstream of the off-ramp (also discussed in Ref. [76]). It can be seen in Figs. 3(b) and 4 that the vehicle speeds slowly decrease within a WSP in the upstream direction whereas the flow rate does not change considerably when a WSP occurs. This is a peculiarity of synchronized flow.

When the vehicle speed in a WSP decreases in the upstream direction, some narrow moving jams emerge in this synchronized flow of low vehicle speed (D17 and D16, Fig. 4). However, D16 is already at the on-ramp (D15-on). For this reason, the synchronized flow propagating upstream, covers this upstream bottleneck at the on-ramp. As a result, an EP occurs [76] (this is not shown in Fig. 4). Thus, WSP upstream of the off-ramp at D25-off and downstream of the on-ramp at D16 is only a part of this EP. Nevertheless, the consideration of this WSP allows us to come to some important conclusions about features of the traffic phase "synchronized flow".

2.3. Overlapping of states of free flow and synchronized flow in density

In particular, when free flow (black quadrates in Fig. 5) and synchronized flow inside WSP (circles in Fig. 5) are shown in the flow–density plane it can be seen, that at least at detectors D20–D18 states of synchronized flow partially overlap with free flow in the density.

The same conclusion can be made if the vehicle speed as a function of the density is drawn (Fig. 6(a)). This means that at the same density either a state of synchronized flow or a state of free flow is possible.

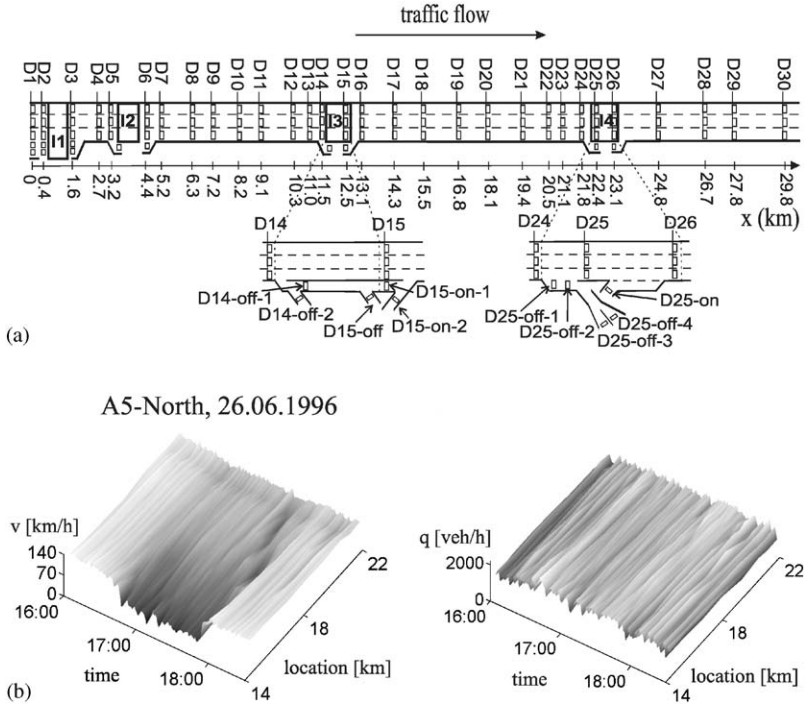

Fig. 3. The widening synchronized flow pattern (WSP): (a) Scheme of the highway infrastructure and local measurements at the section of the highway A5-North [76]; (b) the average speed and the flow rate in WSP as functions of time and location [105].

If now the average absolute values of the vehicle speed difference between the left lane and the middle lane Δv for free flow (curve F in Fig. 6(b)) and for synchronized flow (curve S in Fig. 6(b)) are shown, then we see that there is overlapping of the speed difference in density. This overlapping can lead to a hysteresis between the free flow states and the synchronized flow states. It can be assumed that this overlapping is related to a Z-shaped form of the dependence Δv of the density.[4] Such Z-shaped

[4] The assumption that the overlapping of the speed difference Δv in the density should be related to a Z-shaped characteristic $\Delta v(\rho)$ is made in an analogy to a huge number of physical, chemical and biological systems where states of two different phases overlap in a system parameter. Naturally, there is usually a hysteresis effect due to this overlapping. In the theory of these spatially distributed systems, dependent of the form of this overlapping it is usually assumed that there is one of the N-, S- or Z-shaped characteristics of the system [88–91]. The middle branch of these characteristics cannot usually be observed in experiments. This is because this branch should be related to unstable states of the system [88–91]. The assumption under consideration can be considered as correct if the theory predicts features of spatial–temporal patterns which are observed in experiments [88–91]. The three-phase traffic theory (where the Z-shaped traffic flow characteristic is used [87]) predicts main features of empirical spatial–temporal congested traffic patterns [85–87]. For this reason we will use the assumption about the Z-shaped traffic flow characteristic made above in the further consideration.

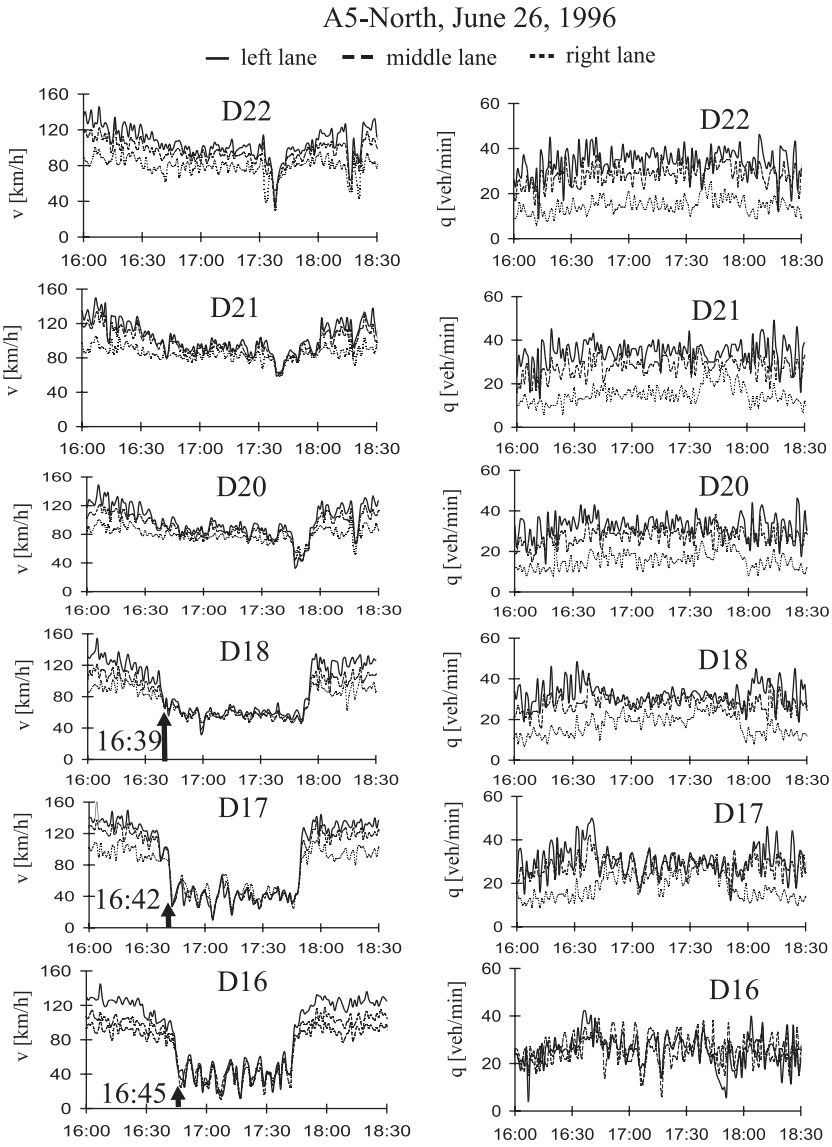

Fig. 4. The widening synchronized flow pattern (WSP): Time series of the vehicle speed (left) and flow rate (right) for different highway lanes for detectors D22–D16. The F → S transitions at the related detectors leading to the WSP formation are marked with up arrows.

characteristic is in agreement with the hypothesis about the Z-shape of the mean probability for passing [12,80,84] (see Section 4.3).

The physical meaning of the result in Fig. 6(b) is the following: In free flow the difference in the average vehicle speed on German highways between the left

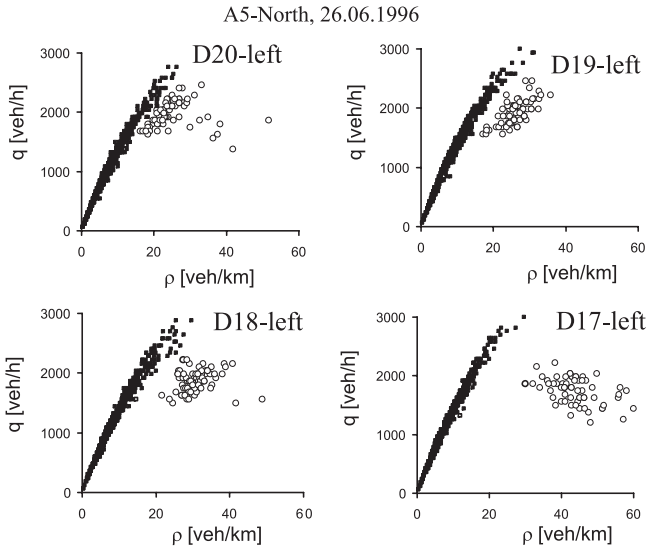

Fig. 5. The measurement points in the flow–density plane for WSP shown in Fig. 3 at the detectors D20–D17 (left line). Free traffic is related to black quadrates, synchronized flow is related to circles. Overlapping of states of free flow and synchronized flow at the detectors D19 is in the density range from about 18 to 36 vehicles/km.

(passing) highway lane and the middle lane due to the high mean probability of passing is considerably higher than that in synchronized flow. However, at the same density in a limited range (e.g., at D19 from 18 vehicles/km to 26 vehicles/km) either states of free flow or synchronized flow can exist. This may lead to a nearly Z-form of the dependency Δv on the vehicle density. The lower the vehicle speed in synchronized flow is the less the density range of the overlapping of the curves F and S (Fig. 6(b), D18). This overlapping disappears fully if the vehicle speed in synchronized flow further decreases.

2.4. Analysis of individual vehicle speeds

To see the difference between free flow and synchronized flow and features of synchronized flow more clearly, distributions of the number of vehicles as a function of the individual vehicle speed for synchronized flow (Fig. 6(c)) and for free flow (Fig. 6(d)) are shown. This is possible because the types of vehicles and their individual vehicle speed during each of the 1 min intervals are also available.

Firstly, it can be seen that in synchronized flow the mean vehicle speed of vehicles and long vehicles are almost the same for different highway lanes whereas for free flow these mean values are strongly shifted to one another.

Secondly, we see that at the detectors D19 during 121 min of the observation individual vehicle speeds in synchronized flow were not lower than 40 km/h (Fig. 6(c), left). At the detectors D18 during 105 min of the observations individual speeds

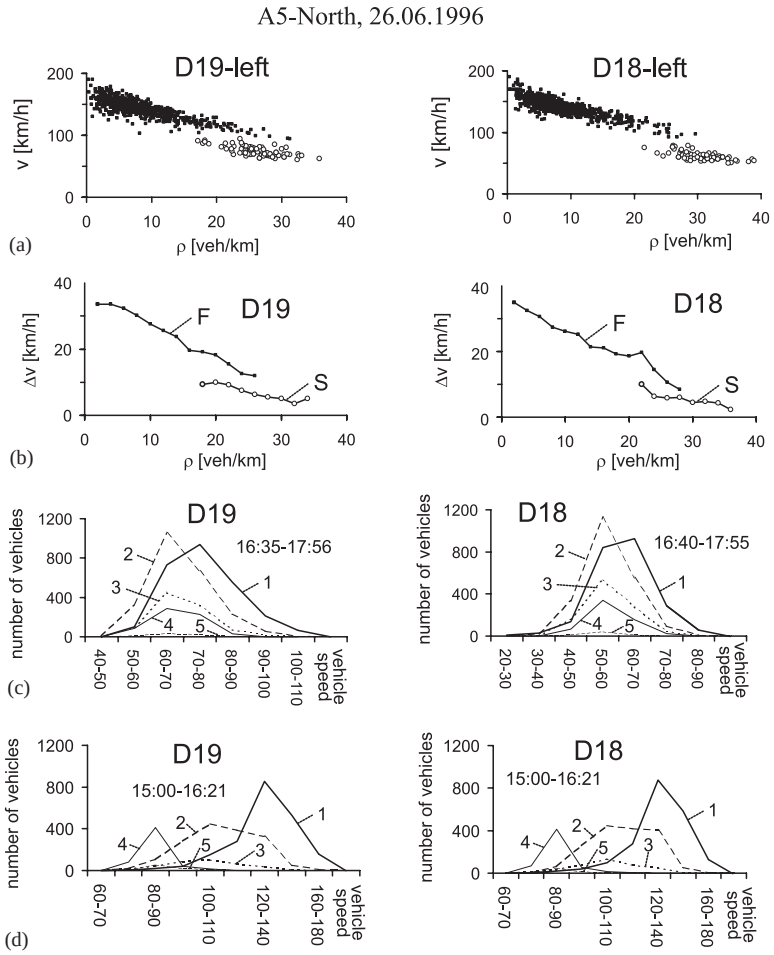

Fig. 6. Empirical features of synchronized flow at the detectors D19 (left) and D18 (right) in the WSP shown in Figs. 3(b) and 4: (a) Measurement points in the speed–density plane (free flow–black quadrates, synchronized flow—circles; left lane). (b) The average difference in the vehicle speeds between left and middle highway lanes as a function of the density, curve *F* for free flow, curve *S* for synchronized flow. The speed differences in (b) are averaged for density intervals of 2 vehicles/km. (c,d) Distribution of the number of vehicles as a function of the different speed classes related to individual single vehicle data for synchronized flow (c) and for free flow (d) The curve 1 is related to vehicles on the left lane, 2—vehicles on the middle lane, 3—vehicles on the right lane, 4—long vehicles on the right lane, 5—long vehicles on the middle lane (long vehicles do not usually move on the left (passing) lane of a three-lane (in one direction) highway in Germany). In (c,d), measured single vehicle data are shown whereby the number of vehicles in each of 15 different classes in regard to the vehicle speed is used separately for vehicles and for long vehicles.

of 6181 vehicles which passed the detectors were measured (Fig. 6(c), right). Among these 6181 vehicles there were no vehicles which had the speed below 20 km/h, there were only 9 vehicles which had individual speeds between 20 and 30 km/h and 59 vehicles which had individual speeds between 30 and 40 km/h. All other

6113 vehicles had individual speeds higher than 40 km/h. Thus, there were no narrow moving jams in synchronized flow between D19 and D18. Nevertheless, these states of synchronized flow cover 2D regions in the flow–density plane (Fig. 5, D19-left, D18-left).

3. Shortcoming of the fundamental diagram approach for description of traffic congestion

Different explanations of empirical features of wide moving jams and synchronized flow [77,74,12,80,79,81] are up to now being discussed between different groups (e.g., Refs. [60–70,97–99,112,71,100,72,85,92–94,102,86] and the reviews [32–34]).

Due to the effort of different scientific groups (see e.g., Refs. [38–40,42–44,46–70,92–94,108] and the reviews [31–34]) considerable progress has been made in the understanding of the theoretical spatial–temporal congested patterns in different traffic flow models within the fundamental diagram approach. In particular, in this approach two main classes of traffic flow models may be distinguished which claim to show moving jams and other congested patterns upstream from an on-ramp:

(i) Models were at a sufficiently high initial flow rate on the main road, upstream from an on-ramp, moving jams spontaneously occur if the flow rate at the on-ramp beginning from zero gradually is increasing. However, the range of the flow rate to the on-ramp where moving jams spontaneously occur is limited. Beginning at a high enough flow rate to the on-ramp spatial homogeneous states of traffic flow which have been called "homogeneous congested traffic" (HCT) [62] occur upstream of the on-ramp where no moving jams spontaneously emerge (e.g., [59,62–65,32]).

(ii) Models were, as well as in the models of item (i), beginning at some flow rate to the on-ramp, moving jams spontaneously occurring upstream of the on-ramp. However, no HCT occurs in these models.

How does the traffic phase "synchronized flow" and the traffic phase "wide moving jam" emerge in an initially free traffic flow at an isolated bottleneck (i.e., the bottleneck is far away from other effective bottlenecks), e.g., at a bottleneck due to an on-ramp? Empirical observations in Refs. [74,76] allow us to conclude that the following scenarios are responsible for the phase transitions and for the pattern evolution in traffic flow at the on-ramp:

(1) Moving jams do *not* emerge in an initial free flow at the on-ramp when the flow rate at the on-ramp is gradually increasing. Rather than moving jams the phase transition from free flow to synchronized flow occurs at the on-ramp.

(2) At a low enough flow rate to the on-ramp the vehicle speed in synchronized flow which has occurred upstream of the on-ramp is relatively high. Moving jams do *not* necessarily emerge in that synchronized flow. If the flow rate to the on-ramp is high, then the vehicle speed in the synchronized flow is low and moving jams, in particular wide moving jams, emerge in that synchronized flow.

(3) The lower the average vehicle speed in synchronized flow upstream of the on-ramp, the higher the frequency of the moving jam emergence in that synchronized flow. This means that the moving jam emergence goes on up to the highest possible values of the flow rate to the on-ramp: Traffic states of high density and low vehicle speed, where moving jams do not emerge, are *not* observed in synchronized flow upstream of the on-ramp.

The empirical results in item (1) and (2) are qualitatively in contradiction with both model classes (i) and (ii) in the fundamental diagram approach. In these models, at high enough initial flow rates on the main road upstream of the on-ramp moving jams must emerge in an initial free flow if the flow rate to the on-ramp beginning from zero is gradually increased [32]. The last empirical result in item 3 means that on average the higher the vehicle density, the lower the stability of traffic flow with respect to the moving jam emergence in that flow. This result of observations which seems to be intuitively obvious for each driver is in a qualitative contradiction with the models of class (i) in the fundamental diagram approach where HCT, i.e., homogeneous congested traffic of high density and low vehicle speed must occur where moving jams do not emerge [32].

The features of the theoretical diagram [62,63,66,32] given a high enough initial flow rate on a highway upstream of the on-ramp may be illustrated with the following simple *theoretical scheme* 1:

- A high flow rate on the main road upstream of the on-ramp *and* a low flow rate to the on-ramp $\rightarrow$ different kinds of moving jams must emerge.
- A high enough flow rate to the on-ramp $\rightarrow$ HCT where the density is high and the speed is very low and no moving jams spontaneously emerge must occur.

In contrast to this theoretical result [62,63,66,32], in empirical observations [76] the following *empirical scheme* 2 is observed:

- A low flow rate to the on-ramp $\rightarrow$ synchronized flow where the density is relatively low and the speed is relatively high occurs where moving jams do not necessarily emerge.
- A high enough flow rate to the on-ramp $\rightarrow$ moving jams must spontaneously emerge in synchronized flow upstream of the on-ramp at any high flow rate to the on-ramp.

Thus, these two schemas (the theoretical scheme 1 [62,63,66,32] and the empirical scheme 2 [76]) are in contradiction with one another.

It should be noted that in models with the fundamental diagram for steady states which should explain empirical spatial–temporal congested traffic patterns beyond the instability of some of these steady states, fluctuations and instabilities let the system evolve in time through a 2D region in the flow–density plane as well. This 2D region is related to some "dynamical" model solutions. Examples of the 2D region in the flow–density plane for the "dynamical" model states are the Nagel–Schreckenberg CA models (e.g., Ref. [69]), the models by Tomer et al. [72], Helbing, Treiber and co-workers [71,73], Nishinari and Takahashi [108], Fukui et al. [94]. However, in these traffic flow models, steady model states are related to a *one-dimensional* region in the flow–density plane, i.e., to the fundamental diagram.

If, in accordance with the three-phase traffic theory, *steady states* of synchronized flow form a 2D region in the flow–density plane, the model dynamics is fundamentally different [85–87]. This leads also to qualitative differences between the patterns of congested traffic obtained in the three-phase traffic theory or in the fundamental diagram approach, respectively.

It must be noted that this critical consideration of the application of the fundamental diagram approach for the description of congested traffic does not concern some important mathematical ideas which have been introduced and developed in models and theories within the fundamental diagram approach with the aim to describe the traffic flow dynamics, for example the ideas about the modelling of vehicle safety conditions, fluctuations, vehicle acceleration and deceleration, different vehicle time delays and other important effects. In particular, the related pioneer mathematical ideas have been introduced in models and theories within the fundamental diagram approach by Lighthill and Whitham [17], Richards [18], Prigogine [23], Gazis et al. [20], Komentani and Sasaki [21], Newell [22], Whitham [45], Bando et al. [47], Payne [24], Gipps [25], Wiedemann [27], Nagel and Schreckenberg and co-workers [42–44,52,69], Takayasu and Takayasu [109], Krauß et al. [51], Mahnke, Kühne et al. [57,107,95], Helbing, Treiber and co-workers [56,66,73], Nishinari and Takahashi [108], Fukui et al. [94,106], Havlin, Tomer and co-workers [68,72], Nagatani and Nakanishi [111] and by many other groups (see references in the reviews by Chowdhury et al. [31], Helbing [32], Nagatani [33], Nagel et al. [34], Wolf [110]). These mathematical ideas are also very important elements of the three-phase traffic theory [12,80,83] and of microscopic models within this theory [85–87]. The main feature of the three-phase traffic theory is that this theory *rejects* the basic hypothesis about the fundamental diagram of earlier traffic flow theories and models. The three-phase traffic theory introduces the new phase of traffic flow, synchronized flow, whose steady states cover a 2D region in the flow density plane. This allows us to overcome the above described problems of the fundamental diagram approach and to explain empirical spatial–temporal congested pattern features [74,76].

4. Comparison of hypotheses to three-phase traffic theory with some results of fundamental diagram approach

4.1. Steady speed states [78,12,80]

4.1.1. Explanation of fundamental hypothesis of the three-phase traffic theory

The fundamental hypothesis of the three-phase traffic theory has already been formulated in Section 1.2.2 [74,78,79,81–83,103,104]: Hypothetical spatially homogeneous and time-independent (stationary) states of synchronized flow, i.e., steady states of synchronized flow where vehicles move at the same distance to one another with the same time-independent vehicle speed cover a 2D region in the flow–density plane (Fig. 1). These steady states are the same for multi-lane and for one-lane roads. In other words, in the three-phase traffic theory there is *no* fundamental diagram for steady states of synchronized flow.

This is not excluded by the empirical fact mentioned above, that a given vehicle density determines the *average* vehicle speed. Indeed, from empirical observations it may be concluded that at the same distance between vehicles (at the same density) there may be a continuum of different vehicle speeds within some finite range in synchronized flow. Obviously the averaging of all these vehicle speeds leads to one average value at the given density.

The 2D region of steady states of synchronized flow (Fig. 1) is also not ruled out by car following experiments, where a driver has the task to follow a specific leading car and not lose contact with it (e.g., Ref. [1]). In such a situation, the gap between the cars will be biased towards the security gap depending on the speed of the leading car. In synchronized flow the situation is different: The gap between cars can be much larger than the security gap.

The hypothesis about steady states of synchronized flow which cover a 2D-region in the flow–density plane makes the three-phase traffic theory [74,78,83] almost *incompatible* with all classical traffic flow theories and present models (see, e.g., Refs. [26,28,29,31,32,38–40,17,20,22,42,45–48,50–52,55–63,66,68–70]) which are based on the fundamental diagram approach.

This follows from the diagram of congested patterns at bottlenecks which has recently been found by the author [76,103,114]. This diagram has been postulated on very general grounds within the three-phase traffic theory [76,103] (see Section 4.6.3) and demonstrated for a microscopic traffic model by Kerner and Klenov [85]. In this model, the upper boundary of a 2D-region of steady states of synchronized flow in the flow–density plane is related to the well-known dependence of the safe speed on the gap between vehicles. The low boundary of these steady states is related to the dependence of a synchronization distance D on the speed. These characteristics of steady states of synchronized flow have been used by Kerner, Klenov and Wolf for the formulation of KKW cellular automata (CA) models to the three-phase traffic theory which may show qualitatively the same diagram of congested patterns [86]. Recently Kerner and Klenov developed a microscopic non-linear theory of spatial–temporal congested patterns at highway bottlenecks [87]. This theory which is based on the three-phase traffic theory allows us to explain and predict main empirical features of congested traffic patterns [74,76].

Note that the three-phase traffic theory is *a behavioral theory* of traffic flow. This means that the hypotheses of this theory are based on common behavioral fundamental characteristics of drivers observed on highways. In particular, the fundamental hypothesis of the three-phase traffic theory (Section 2) is linked to a driver's ability to recognize whether the distance to the vehicle ahead becomes higher or lower over time [78]. If gaps between vehicles are not very high, this driver's ability is true even if the difference between vehicle speeds is negligible.

It has been shown [76,85–87] that the main *qualitative* features of the congested pattern emergence observed in empirical investigations [74,76] can be shown in the three-phase traffic theory where all drivers have the same characteristics and all vehicles have the same parameters. Obviously, in real traffic there are differences in driver's characteristics and in vehicle parameters (e.g., different desired speeds and different safe speeds, aggressive and timid driver's behavior, vehicles and long vehicles) which may

change some spatial–temporal congested pattern parameters and conditions of the pattern emergence. However, these differences in driver's characteristics and in vehicle parameters may first be neglected when *fundamental* congested traffic pattern features are studied. This statement is confirmed in the microscopic theory of congested traffic patterns [87] which is able to explain and to predict main fundamental empirical congested pattern features.

4.1.2. Hypothesis about stability of steady states [12,78,80]

In 1959 Herman, Montroll, Potts and Rothery have introduced the concept of a driver's acceleration (deceleration) delay time in their car following microscopic traffic flow model. They found an instability in some of the steady states of traffic flow [19]. The existence of the delay times in the vehicle acceleration and deceleration has been confirmed in a lot of empirical observations (e.g., Ref. [26]). To study traffic phenomena beyond the instability, in 1961 Gazis, Herman and Rothery [20] and Newell [22] used the fundamental diagram for steady model states (for the review see Ref. [34]). Later the driver's delay time and the fundamental diagram for steady model states have been used in a huge number of other traffic flow models and theories where the linear instability of steady states should occurs when the vehicle density exceeds some *critical* vehicle density (see references in the reviews [31–34]).

In the three-phase traffic theory, the driver's delay time plays a very important role also, however a different hypothesis is suggested [78,83,74,80,12,84]: *Independent of the vehicle density in a steady state of synchronized flow infinitesimal perturbations of any traffic flow variables* (*e.g.*, *the vehicle speed and/or the gap*) *do not grow*: In the whole possible density range steady states of synchronized flow can exist. In other words, in the whole possible density range (Fig. 1) there are no unstable steady states of synchronized flow with respect to infinitesimal perturbations of any traffic flow variables.

To explain this hypothesis from the driver's behavior, let us consider a small enough fluctuation in the braking of a vehicle in an initial synchronized flow steady state which is related to one of the steady states inside a 2D region in the flow–density plane (dashed region in Fig. 1). This small braking may lead to a transition from the initial synchronized flow state to another state with a lower gap. This lower gap is equal to a gap in another state in a 2D region in the flow–density plane. Thus, an occurrence of this fluctuation may cause a spatial–temporal transition to another state of synchronized flow. Therefore drivers should not immediately react on this transition. For this reason even after a time delay, which is due to a finite reaction time of drivers, the drivers upstream should not brake stronger than drivers in front of them to avoid an accident. As a result, a local perturbation of traffic variables (density or vehicle speed) of small enough amplitude does not grow.

Already small amplitude fluctuations make real traffic flow always non-homogeneous and non-stationary. Thus, steady states are only hypothetical traffic flow states which cannot probably be found in measurements of real traffic flow on a freeway. The measurements of real traffic flow in which a broad spread of vehicle gaps at the same speed is observed (e.g. [1]) may be only a hint (but it is obviously not a proof) for the fundamental hypothesis about a 2D region of the steady states of synchronized flow in

the three-phase traffic theory. As a proof of the three-phase traffic theory mathematical results of this theory can be considered [85–87]. These results allow us to overcome the problems of the fundamental diagram approach for the description of empirical features of the phase transitions and spatial–temporal congested patterns.

4.2. Phase transitions on homogeneous (without bottlenecks) roads

4.2.1. The line J and hypothesis about emergence of moving jams [74]

In contrast to the traffic flow theories in the fundamental diagram approach [31–34], in the three-phase traffic theory, there is *no* critical density where traffic flow should become unstable with respect to the moving jam emergence [74,83,78]. How is the moving jam emergence explained in the three-phase traffic theory?

To understand this, we consider first the process of the driver's escaping from a standstill inside a wide moving jam. This process determines the velocity of the upstream motion of the downstream front of the wide moving jam. In this process, each driver standing inside the wide moving jam can start to escape from the jam after (i) the vehicle in front of the driver has already escaped from the jam and (ii) the distance between these vehicles has exceeded some "safety distance". Thus, the driver begins to escape after a delay time in the vehicle acceleration. The related mean delay time $\tau_{\rm del}^{\rm (a)}$ determines the average time interval between two vehicles following one another escaping from the wide moving jam. Therefore, the velocity of the downstream jam front $v_{\rm g}$ is:

$$v_{\rm g} = -\frac{1}{\rho_{\rm max}\tau_{\rm del}^{\rm (a)}}\,, \tag{1}$$

where $\rho_{\rm max}$ is the mean vehicle density inside the wide moving jam.

We propose that all vehicles are the same and all drivers have the same characteristics. Thus, we can suggest that the mean parameters $\rho_{\rm max}$ and $\tau_{\rm del}^{\rm (a)}$ do not depend on time. This means that the velocity of the downstream wide moving jam front $v_{\rm g}$ (1) is also independent of time: The propagation of the downstream wide moving jam front is on average a stationary process. The stationary motion of the downstream jam front can be presented in the flow–density plane by the characteristic line J (Fig. 7(a)) which is called the line J [77]. The slope of the line J in the flow–density plane is equal to the velocity of this downstream front $v_{\rm g}$. If free flow is formed in the wide moving jam outflow, the flow rate in this jam outflow is $q_{\rm out}$, the density is $\rho_{\rm min}$ and the speed is $v_{\rm max}$. As well as the velocity $v_{\rm g}$, the mean parameters of the jam outflow $q_{\rm out}$, $\rho_{\rm min}$ and $v_{\rm max}$ are characteristic parameters which do not depend on initial conditions. The stationary propagation of the downstream front of wide moving jams and the line J have been found in empirical studies of wide moving jams [77,113].

The following hypothesis of the three-phase traffic theory is related to the wide moving jam emergence in the 2D steady states of synchronized flow [83,74]:

All (an infinite number!) *steady states of traffic flow which are related to the line J in the flow–density plane are* threshold states *with respect to the wide moving jam emergence.*

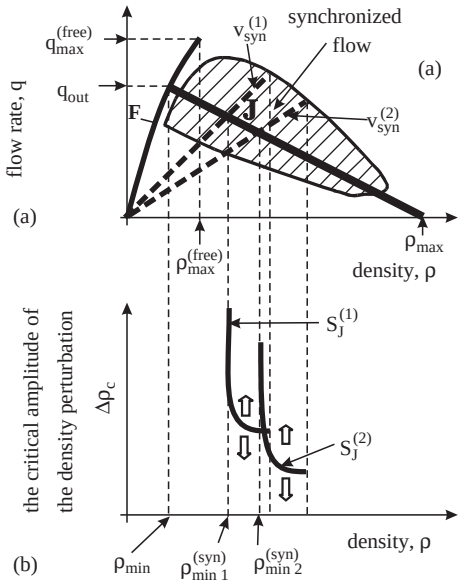

Fig. 7. Explanation of the jam emergence in synchronized flow in the three-phase traffic theory [71]: (a) the line J. States of free (curve F) and steady states of synchronized flow (hatched region) are the same as in Fig. 1(a). (b) Qualitative dependences $(S_J^{(1)}, S_J^{(2)})$ of the critical density amplitude of a local perturbation $\Delta\rho_c$ on the density in steady speed states. The curves $S_J^{(1)}(\rho)$ and $S_J^{(2)}(\rho)$ are related to two different constant speeds $v_{syn}^{(1)}$ and $v_{syn}^{(2)}$, respectively ($v_{syn}^{(1)} > v_{syn}^{(2)}$). The densities $\rho_{min\,1}^{(syn)}$ and $\rho_{min\,2}^{(syn)}$ are the threshold densities for the jam emergence for the vehicle speeds $v_{syn}^{(1)}$ and $v_{syn}^{(2)}$, respectively.

This means that the line J separates all steady states of traffic flow into two qualitatively different classes:

(1) In states which are related to points in the flow–density plane lying below (see axes in Fig. 7(a)) the line J *no* wide moving jams can either continue to exist or be excited.
(2) States which are related to points in the flow–density plane lying on and above the line J are *metastable steady states* with respect to the wide moving jam emergence (the $F \rightarrow J$ transition) where the related nucleation effect can be realized. In a metastable steady state, a growth of an initial local perturbation can lead to the wide moving jam formation *only if* the amplitude of this perturbation exceeds a critical amplitude. In contrast, if the amplitude of the perturbation is lower than the critical amplitude the initial perturbation does not lead to the wide moving jam formation. In the latter case, a transition from one to another steady state in the 2D region of steady states is however possible rather than the wide moving jam formation.

This hypothesis has been confirmed by numerical simulations of the model by Kerner and Klenov in the frame of the three-phase traffic theory (see Fig. 1(f) in Ref. [85]).

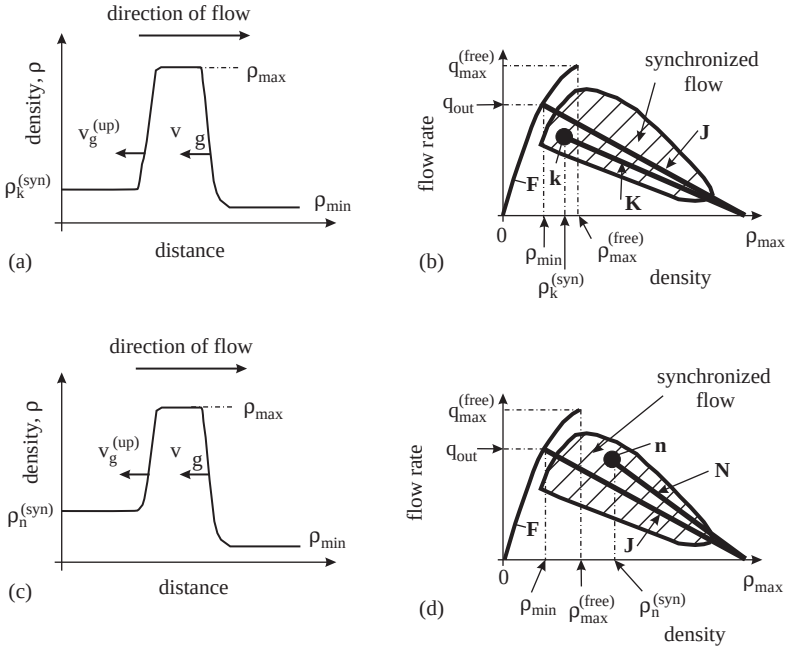

Fig. 8. (a–d) Explanation of the wide moving jam emergence in steady states of traffic flow in Ref. [71]. In (a,c) $q_{\rm out}$ $\rho_{\rm min}$ are the flow rate and the density in the wide moving jam outflow, respectively (see Fig. 7).

To explain the hypothesis, note that wide moving jams cannot be formed in any steady states of traffic flow situated below the line J. Indeed, let the steady state of flow directly upstream of a wide moving jam is related to a point k in the flow–density plane which is below the line J (Fig. 8(a,b)). Because the velocity of the upstream front of the wide moving jam $v_{\rm g}^{\rm (up)}$ equals the slope of a line from k to the point $[\rho_{\rm max}, 0]$, the related absolute value $|v_{\rm g}^{\rm (up)}|$ is always lower than that of the downstream front $|v_{\rm g}|$ which is determined by the slope of the line J:

$$|v_{\rm g}^{\rm (up)}| < |v_{\rm g}| \ . \tag{2}$$

Therefore, the width of the wide moving jam is gradually decreasing. Otherwise, let us consider the case when a steady flow which is upstream from another wide moving jam is above the line J (see a point n in Fig. 8(d)). In this case, the velocity of the upstream front of the wide moving jam $v_{\rm g}^{\rm (up)}$ equals the slope of a line from n to the point $[\rho_{\rm max}, 0]$, i.e., the related absolute value $|v_{\rm g}^{\rm (up)}|$ is always higher than that of the downstream front $|v_{\rm g}|$:

$$|v_{\rm g}^{\rm (up)}| > |v_{\rm g}| \ . \tag{3}$$

Therefore, the width of the wide moving jam in Fig. 8(d) should be gradually increasing. For these reasons, wide moving jams can be formed in steady states of traffic flow which lie on and above the line J.

Recall that the line J has first been introduced by Kerner and Konhäuser in 1994 [46] in their theory of wide moving jams. The Kerner–Konhäuser theory of wide moving jams which has been derived within the fundamental diagram approach has been further applied and developed for a number of different traffic flow models [48,51,52,56,57,62,63,66,69,70,31,32,95]. In this theory [46], the line J separates stable and metastable states of free flow with respect to the wide moving jam emergence: States of free flow which are on and above the line J (i.e., when the density and the flow rate in free flow are either equal to or higher than the density $\rho_{\min}$ and the flow rate q_{out} in the outflow from the wide moving jam when free flow is formed downstream of the jam, respectively) are metastable states with respect to the wide moving jam emergence (the $\mathrm{F} \to \mathrm{J}$ transition). The line J and the characteristic wide moving jam parameters following from the theory [46] have indeed been found in empirical observations [113].

However, in the fundamental diagram approach at the limit point of this metastable region of free flow $(\rho_{\max}^{(\text{free})}, q_{\max}^{(\text{free})})$ (Fig. 7(a)) free flow becomes unstable with respect to the moving jam emergence (the $\mathrm{F} \to \mathrm{J}$ transition) [46–48,50,55,56,58–63,66,68 31–34,95]. This common feature of all traffic flow models within the fundamental diagram approach which claim to show moving jam emergence is in contradiction with empirical observations: Moving jams do not spontaneously emerge in real free flow. In contrast, in the three-phase traffic theory at this limit point of free flow the $\mathrm{F} \to \mathrm{S}$ transition occurs rather than the $\mathrm{F} \to \mathrm{J}$ transition [12,80]. This is in accordance with empirical results [74,76] (see for more detail Section 4.2.5).

4.2.2. Hypothesis about continuous spatial–temporal transitions between states of synchronized flow on homogeneous roads [78]

If in a metastable steady state on and above the line J in the flow–density plane a local perturbation occurs whose amplitude is lower than the critical amplitude then this perturbation does not lead to the moving jam emergence. However, the perturbation can nevertheless cause a local transition to another state of synchronized flow rather than traffic flow returning to the initial steady state. The same case can also occur for steady states below the line J, i.e., for steady states which are stable with respect to the moving jam emergence. Thus, when a lot of different local perturbations appear in traffic flow, then very complex spatial–temporal transitions between states of synchronized flow can occur. These synchronized flow states can be close to steady states. This is the contents of the following hypothesis of the three-phase traffic theory [78,83,74,80,12,84]:

Local random perturbations in synchronized flow can cause continuous spatial–temporal transitions between different states of synchronized flow (a random "walking" between different 2D states in the flow density plane).

Complex transformation between small-amplitude spatial–temporal states which are close to model steady states is observed in microscopic models [85,86] where steady speed states of synchronized flow cover a 2D-region in the flow density plane.

Strictly speaking, already small-amplitude random perturbations and deterministic dynamical effects in synchronized flow destroy steady states: Rather than steady states some complex spatial–temporal dynamical states of synchronized flow can appear. However, if the perturbation amplitude is low enough these complex states can be

very close to steady states. These small-amplitude spatial–temporal states should possess the features of steady states discussed above and which will be considered in other hypotheses of the three-phase traffic theory below. This behavior is observed in the recent models based on the three-phase traffic theory [85–87].

4.2.3. Two kinds of nucleation effects and phase transitions in free flow

How should the traffic phase "synchronized flow" and the traffic phase "wide moving jam" emerge in an initial free traffic flow on a homogeneous multi-lane road (curve *F* in Fig. 1(a))? The following hypothesis of the three-phase traffic theory answers these questions (Figs. 9 and 11) [80,12]:

At the same density in free flow there may be two qualitatively different nucleation effects and the related two qualitative different first-order phase transitions in free flow:

(1) The nucleation effect which is responsible for the moving jam's emergence in free flow, i.e., for the phase transition from free flow to a wide moving jam (the F → J transition).
(2) The nucleation effect which is responsible for the F → S transition.

The F → S transition is observed in real free flow: This transition is responsible for the onset of congestion in free flow [76]. The onset of congestion is accompanied by the speed breakdown. This is called *the breakdown phenomenon* in free flow (e.g., Ref. [8]). Thus, in the three-phase traffic theory, the well-known breakdown phenomenon (the onset of congestion) is explained by the F → S transition [83,12]. The empirical breakdown phenomenon, i.e., the F → S transition is accompanied by a hysteresis effect. The hysteresis effect is an attribute of the first order phase transition.

Although the spontaneous F → J transition is not observed, nevertheless the F → J transition can be *induced* in free flow. This is confirmed by two empirical facts: (i) The induced F → J transition has been observed in Ref. [81]. (ii) Empirical maximum flow rate in free flow $q_{\max}^{(\mathrm{free})}$ satisfies the condition:

$$q_{\max}^{(\mathrm{free})} > q_{\mathrm{out}} \,. \tag{4}$$

This means that there are states of free flow where the flow rate $q > q_{\mathrm{out}}$. These states are metastable with respect to the wide moving jam emergence.

Thus, both phase transitions are confirmed by empirical data. These both first-order phase transitions are also found in a microscopic three-phase traffic theory [85–87].

4.2.4. Priority of F → S transition in free flow [12,80]

The following hypothesis answers the question whether the F → J transition or the F → S transition is more probable in free flow at the same density:

At each given density in free flow the critical amplitude of a local perturbation in the free flow which is needed for the F → S transition (the curve F_{S} in Fig. 9(b)) is considerably lower than the critical amplitude of a local perturbation which is needed for the F → J transition (the curve F_{J} in Fig. 9(b)).

This hypothesis is confirmed by empirical observations: The onset of congestion in real free flow is linked to the F → S transition rather than to the F → J transition

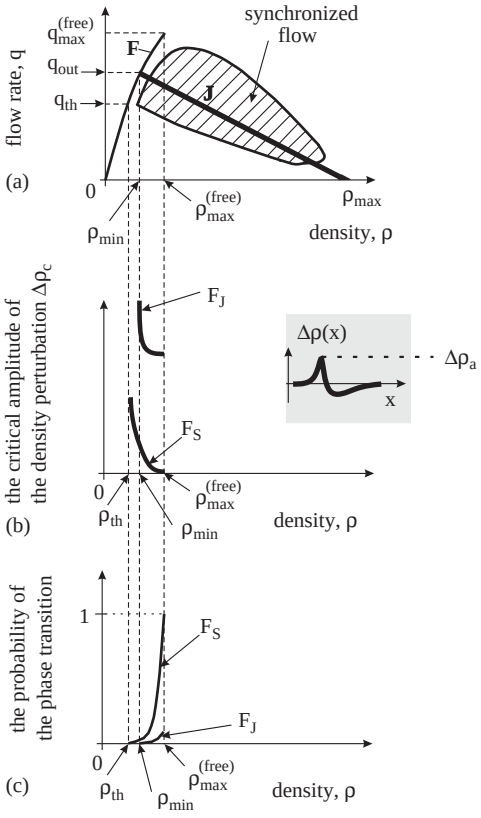

Fig. 9. Explanation of hypotheses to the three-phase traffic theory [12,80]: (a) States of free (curve F) and synchronized flow (hatched region) which are the same as in Fig. 1(a). (b) Qualitative dependencies of the amplitude of the critical density local perturbation on the density. (c) Qualitative dependencies of the probability of the phase transitions. In (b,c) the curve F_S is related to the $F \to S$ transition and the curve F_J is related to the $F \to J$ transition. In (b), right, a form of a local density perturbation is schematically shown which grows if the amplitude of this perturbation, $\Delta\rho_a$, exceeds the critical amplitude $\Delta\rho_c$ (links). Qualitative character of dependencies of the critical amplitude of the speed local perturbation on the density is the same as for the density perturbation shown in (b). This is because a change in the density inside a local perturbation in free flow is accompanied by the related change in the speed.

[74,76]. This is also a result of a mathematical three-phase traffic theory of the phase transitions in free flow (see Fig. 1(b) in Ref. [85]).

Obviously, the higher the amplitude of a random local perturbation in a free flow, the lower the mean probability (for a given time interval) of an occurrence of this perturbation. Thus, at each given density in free flow the mean probability for a given time interval of an occurrence of the $F \to S$ transition, P_{FS} (the curve F_S in Fig. 9(c)) should be considerably higher than the probability for the same time interval of the $F \to J$ transition, P_{FJ} (the curve F_J in Fig. 9(c)).

4.2.5. The critical (limit) density in free flow [12,80]

The following hypothesis explains the existence of the critical (limit) point of free flow $(\rho_{\max}^{(\mathrm{free})}, q_{\max}^{(\mathrm{free})})$ (Figs. 1(a) and 9(a)):

The existence of the critical (limit) point of free flow $(\rho_{\max}^{(\mathrm{free})}, q_{\max}^{(\mathrm{free})})$ *is linked to the* $F \to S$ *transition rather than with the moving jam emergence.* This hypothesis is related to the empirical result that in each free flow state, if the onset of congestion in this free flow occurs then this is linked to the $\mathrm{F} \to \mathrm{S}$ transition rather than to the $\mathrm{F} \to \mathrm{J}$ transition [74,76]. The following hypotheses also result from this empirical fact:

At the limit point of free flow $(\rho_{\max}^{(\mathrm{free})}, q_{\max}^{(\mathrm{free})})$ *the probability of the* $F \to S$ *transition,* P_{FS} *reaches one and the critical amplitude of the local perturbation for the* $F \to S$ *transition reaches zero* (the curves F_{S} in Fig. 9(b,c)).

At the limit point of free flow $(\rho_{\max}^{(\mathrm{free})}, q_{\max}^{(\mathrm{free})})$ *the probability of the* $\mathrm{F} \to \mathrm{J}$ *transition is very low and consequently the critical amplitude of a local perturbation which is needed for the* $\mathrm{F} \to \mathrm{J}$ *transition is a relatively high finite value* (the curves F_{J} in Fig. 9(b,c)).

These hypotheses are also confirmed by a microscopic three-phase traffic theory [85–87] where the limit point of free flow is linked to the $\mathrm{F} \to \mathrm{S}$ transition rather than to the $\mathrm{F} \to \mathrm{J}$ transition.

In contrast to these results of the three-phase traffic theory and to results of empirical observations [76], in the fundamental diagram approach the breakdown phenomenon as well as the limit point of free flow $(\rho_{\max}^{(\mathrm{free})}, q_{\max}^{(\mathrm{free})})$ are explained by the moving jam emergence in free flow (see references in the reviews [31–34]).

4.3. Physics of breakdown phenomenon

4.3.1. Z-shaped speed–density characteristic

Because corresponding to the hypothesis discussed in Sections 4.2.3–4.2.5 the empirical onset of congestion (the breakdown phenomenon) in an initial free flow is linked to the $\mathrm{F} \to \mathrm{S}$ transition, we consider here the physics of this phase transition in more detail. To do this, let us designate the critical amplitude of the critical local perturbation needed for the $\mathrm{F} \to \mathrm{S}$ transition as $\Delta v_{\mathrm{cr}}^{(\mathrm{FS})}$. For this critical amplitude we can write the formula

$$\Delta v_{\mathrm{cr}}^{(\mathrm{FS})} = v_{\mathrm{initial}}^{(\mathrm{free})} - v_{\mathrm{cr}}^{(\mathrm{FS})} , \tag{5}$$

where $v_{\mathrm{initial}}^{(\mathrm{free})}$ is the speed in an initial free flow and $v_{\mathrm{cr}}^{(\mathrm{FS})}$ is the vehicle speed inside the critical local perturbation. At the critical (limit) point of free flow $(\rho_{\max}^{(\mathrm{free})}, q_{\max}^{(\mathrm{free})})$ the critical amplitude

$$\Delta v_{\mathrm{cr}}^{(\mathrm{FS})}\big|_{\rho=\rho_{\max}^{(\mathrm{free})}} = 0 . \tag{6}$$

This means that at the limit point of free flow the mean probability for the $\mathrm{F} \to \mathrm{S}$ transition, P_{FS} is equal to

$$P_{\mathrm{FS}}\big|_{\rho=\rho_{\max}^{(\mathrm{free})}} = 1 . \tag{7}$$

The critical (limit) density $\rho_{\max}^{(\mathrm{free})}$ (the critical flow rate $q_{\max}^{(\mathrm{free})}$) is found from condition (7). Thus, the critical (limit) point of free flow $(\rho_{\max}^{(\mathrm{free})}, q_{\max}^{(\mathrm{free})})$ is determined by

formula (7), i.e., through the calculation of the mean probability for the F → S transition, P_{FS}. For this reason, we consider the definition of this probability.

This probability is defined as follows. We consider a large number of different realizations, N_{FS}, where the F → S transition in an initial free flow are studied. Each of the realizations should be performed at the same flow rate of free flow, other initial conditions, and during the same time interval T_{ob} of the observation of the spontaneous F → S transition on a chosen highway section of the length L_{ob}. Let us assume that in n_{FS} of these N_{FS} realizations the F → S transition occurs. Then, the mean probability of this transition for the time interval T_{ob} and for the length of the road L_{ob} is equal to [5]

$$P_{FS} = \frac{n_{FS}}{N_{FS}} . \tag{8}$$

When the density in free flow decreases, the critical amplitude of the local perturbation for the F → S transition increases. There should be a threshold point (ρ_{th}, q_{th}) for the F → S transition. This threshold point is defined as follows. If the density decreases then the threshold density ρ_{th} is the *minimum* density in free flow where the F → S transition can *still* occur. This means that below the threshold density, i.e., when $\rho < \rho_{th}$, for the mean probability P_{FS} we have the condition:

$$P_{FS}|_{\rho<\rho_{th}} = 0 . \tag{9}$$

In the threshold point $\rho = \rho_{th}$ $(q = q_{th})$, the critical amplitude of the local perturbation for the F → S transition reaches a maximum value

$$\Delta v_{cr}^{(FS)}|_{\rho=\rho_{th}} = \Delta v_{cr,max}^{(FS)} . \tag{10}$$

This behavior of the critical amplitude of the local perturbation and of the mean probability of the F → S transition P_{FS} is shown by the curves F_S in Fig. 9(b,c), respectively.

The curve F for states of free flow, the critical branch $v_{cr}^{(FS)}(\rho)$ (Fig. 10) which determines the critical amplitude (5) of local perturbation for the F → S transition as a function of the density, and the 2D region for steady states of synchronized flow form together a Z-shaped function of the speed of the density (Fig. 10(b,c)). This hypothesis about the Z-shaped traffic flow characteristic for the breakdown phenomenon is confirmed in the microscopic three-phase traffic flow theory (see Fig. 2(a) in Ref. [87]).

4.3.2. Passing probability

This Z-shaped speed–density relationship is qualitatively correlated with the hypothesis about the mean probability of passing P on a multi-lane road (Fig. 11). In the three-phase traffic theory, the following hypothesis is valid [12,80,84]:

The mean probability of passing, P, on a multi-lane road is a Z-shaped function of the density (Fig. 11(b)).

At a very low vehicle density in free flow vehicles can freely pass. Thus, the probability of passing P in free flow should reach one at the density $\rho \to 0$ (the curve P_F in Fig. 11(b)). To determine the probability of passing P, a large number, N_P, of different realizations (runs) for passing should be performed at the same initial conditions

[5] Strictly speaking the exact mean probability of the spontaneous F → S transition is determined when $N_{FS} \to \infty$.

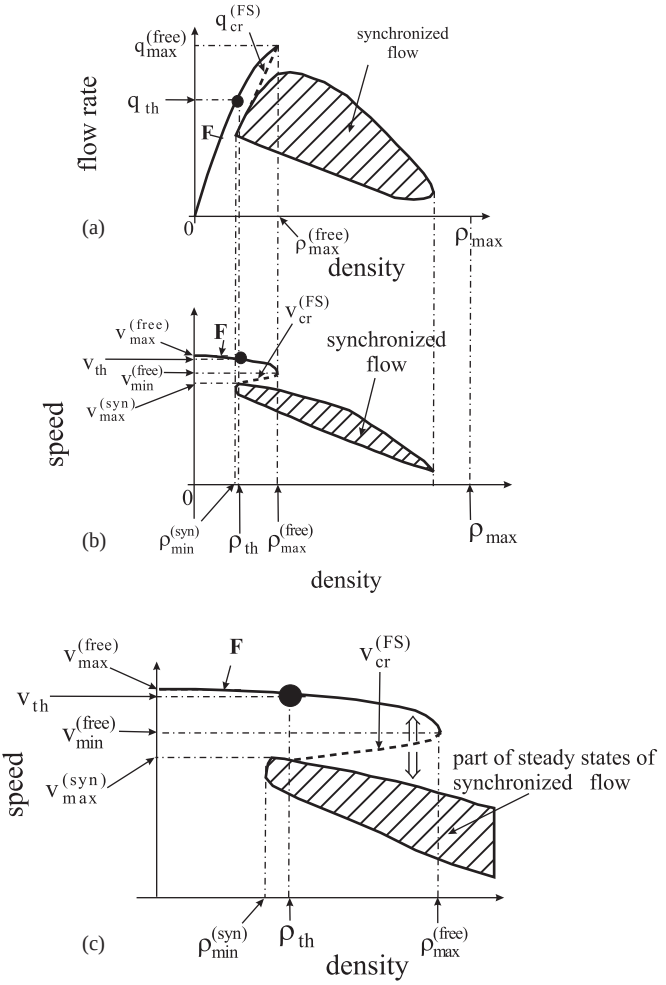

Fig. 10. Qualitative illustration of the F → S transition (the breakdown phenomenon): The critical branch $v_{cr}^{(FS)}$ (dashed curve) gives the speed inside the critical local perturbation, the threshold density ρ_{th} and the threshold flow rate q_{th} in free flow for the F → S transition on a homogeneous road in the flow–density plane (a) and in the speed–density plane (b,c). In (a) states of free and synchronized flows are taken from Fig. 1(a). In (b) states of free and synchronized flows are related to (a). In (c) a part of (b) for lower density range in a higher scale is shown. The black point in (a–c) on the curve F for free flow shows the threshold point $\rho = \rho_{th}$ for the F → S transition.

and the same time interval for passing. In each of these realizations, there should be a driver who moves with a speed higher than the speed of the vehicle ahead. When approaching the vehicle ahead, the driver should try to pass using a passing freeway lane. It can occur that in some of the realizations the driver is able to pass, but in the other realization the driver is not able to pass. The latter is because the passing lane was occupied by other drivers. If the number of realizations where the driver was able

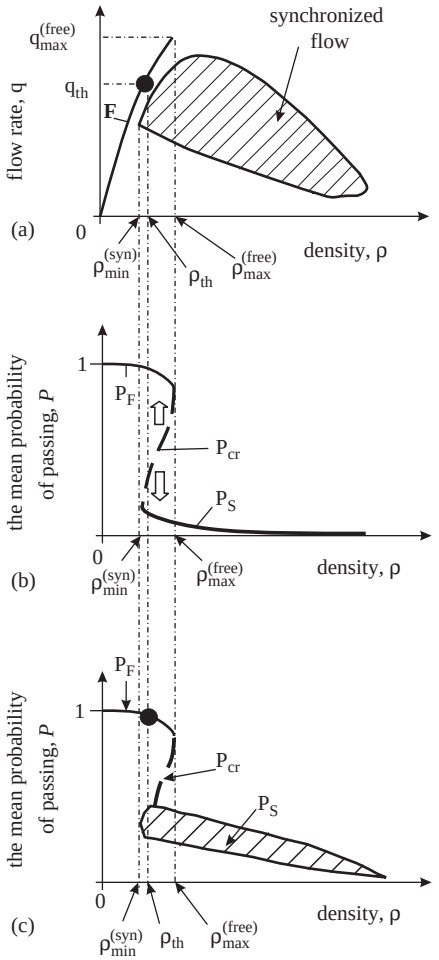

Fig. 11. Explanation of the hypothesis about the Z-shaped dependence of the mean probability of the passing P in the three-phase traffic theory [12,80]: (a) States of free (curve F) and synchronized flow (hatched region) which are the same as in Fig. 1(a). (b) Qualitative dependence of the mean probability of the passing P (which is averaged over all different steady states of synchronized flow at a given density) as a function of the density. (c) Qualitative dependence of the mean probability of the passing P as a function of the density without averaging of the mean probability for all different steady states of synchronized flow at a given density. The black point in (a, c) is related to the threshold point for the $F \to S$ transition.

to pass is n_P, then the probability of passing is [6]

$$P = \frac{n_P}{N_P} . \tag{11}$$

To explain the hypothesis about the Z-shaped form of the probability of passing, note that in accordance with the fundamental hypothesis of the three-phase traffic theory

[6] Strictly speaking the exact probability of passing is determined when $N_P \to \infty$.

steady states of synchronized flow overlap states of free flow in the vehicle density within a density range (Fig. 11(a))

$$[\rho_{\min}^{(\mathrm{syn})}, \rho_{\max}^{(\mathrm{free})}] \,. \tag{12}$$

However, it is well-known that the mean probability of passing in congested traffic (in our case, in one of the two traffic phases in congested traffic, "synchronized flow") P is considerably lower than in free flow. Thus, in free flow the mean probability of passing, P (the curve P_{F} in Fig. 11(b)) should be higher than that in synchronized flow (the curve P_{S}). In the range of the density (12), at the same given density there can be either a state of free flow where $P = P_{\mathrm{F}}$ is high or a steady state of synchronized flow where $P = P_{\mathrm{S}}$ is low. This leads to the Z-shaped function of the mean probability of passing in the three-phase traffic theory.

The Z-form of the dependence of the mean probability of passing P of the density as well as the related Z-shape of the speed–density relationship allows us to explain the $\mathrm{F} \to \mathrm{S}$ transition on a multi-lane (in one direction) freeway. If the density in free flow is gradually increasing then a drop in the mean probability of passing P must occur when the density reaches the limit density for free flow $\rho_{\max}^{(\mathrm{free})}$ (Fig. 11(b)). As a result of this drop the mean probability of passing P decreases sharply to low values of P which are related to synchronized flow. Thus, at the limit density for free flow $\rho_{\max}^{(\mathrm{free})}$ the spontaneous $\mathrm{F} \to \mathrm{S}$-transition must occur.

The hypothesis about the Z-shaped mean probability of passing P as a function of the vehicle density can also be explained by the above empirical study of the WSP where it has been shown that states of synchronized flow overlap with states of free flow in the density (Figs. 5 and 6(a); see also Fig. 2 in Ref. [80]). This means that at the same density either a state of synchronized flow or a state of free flow is possible. It is obvious that the mean probability of passing is higher in free flow than in synchronized flow. Thus, the empirical fact that states of free flow and synchronized flow overlap in the vehicle density (Figs. 5 and 6(a)) means that the mean probability of passing should have a Z-shape. Indeed, the lower the mean rate of passing, the lower the mean probability of passing P is. The mean rate of the passing should decrease when the vehicle speed difference Δv becomes lower.

However, it must be noted that this overlapping of states of free flow and states of synchronized flow occurs in a narrow range of the density in the vicinity of the limit density in free flow $\rho_{\max}^{(\mathrm{free})}$, i.e., when the average speed of synchronized flow is relatively high (see Fig. 2 in Ref. [80] and Section 2.2). Therefore, if in some empirical data *only* synchronized flows of a relatively low vehicle speed are observed then no overlapping of states of free and synchronized flows could be found. As a result, some authors make a conclusion that there is no overlapping of states of free and synchronized flows, and as a result a dependence of the mean probability of passing P as a function of the density should be a monotonous decreasing one (e.g., Ref. [102]). However, to make the correct conclusion about of whether an overlapping of states of free and synchronized flows exists, a more precise empirical study is necessary (Section 2.2).

4.3.3. Competition between over-acceleration and speed adaptation

The first-order $F \to S$ transition may be explained by a *competition* between two contradicting tendencies inside a local perturbation in an initial free flow:

(i) A tendency to the initial free flow due to an "over-acceleration".
(ii) A tendency to a synchronized flow due to the vehicle speed adaptation to the speed of the vehicle ahead.

This hypothesis is confirmed by the microscopic three-phase traffic theory [87].

The vehicle "over-acceleration" can occur due to passing: For passing the vehicle has usually to increase its speed. The tendency to synchronized flow can occur due to the need in the adaptation of the vehicle speed to the speed of the leading vehicle when passing is not possible or is difficult.

The "over-acceleration" is stronger at a higher vehicle speed, exactly lower density. In this case, an initial local perturbation in free flow decays (up-arrow in Fig. 10(c)). The over-acceleration is also responsible for the phase transition from synchronized flow to free flow (the $S \to F$ transition).

In contrast, the tendency to the speed adaptation is stronger at a lower speed, i.e., at a higher density. This causes a decrease in the average vehicle speed inside the initial perturbation (down-arrow in Fig. 10(c)) and a self-maintenance of the emerging synchronized flow.

The curve P_S in Fig. 11(b) is related to some simplifications: These curves are an averaging of all different synchronized flow speeds at a given density to one average speed. If we consider all these different steady-state synchronized flow speeds (as it has been made in Fig. 10(b,c)) we naturally come to Fig. 11(c) for the mean probability of passing which reflects the 2D region of steady states of synchronized flow.

4.4. Why moving jams do not emerge in free flow

In the three-phase traffic theory, wide moving jams do not spontaneously emerge in free flow (Section 4.2.3). This hypothesis is related to results of empirical observations [74,81] and it is also confirmed in the mathematical three-phase traffic theory [85–87].

To explain this hypothesis of the three-phase traffic theory, we present the line J in the flow–density plane together with states of free flow and steady states of synchronized flow as is now shown in Fig. 12(a). The line J in the flow–density plane (Fig. 12(a)) is related to a parabolic function in the speed-density plane (curve J in Fig. 12(b)).

We suggest that besides the critical point of the $F \to S$ transition, $(\rho_{max}^{(free)}, q_{max}^{(free)})$, there is another hypothetical critical point in free flow, $(\rho_{max,\ FJ}^{(free)}, q_{max,\ FJ}^{(free)})$, where the critical amplitude of the critical perturbation which is needed for the $F \to J$ transition is zero (Fig. 12(b)). However, the density in free flow which is related to this critical density $\rho_{max,\ FJ}^{(free)}$ is *much higher* than the critical density $\rho_{max}^{(free)}$ where the $F \to S$ transition must occur. This hypothesis is confirmed by the microscopic three-phase traffic theory [85,87].

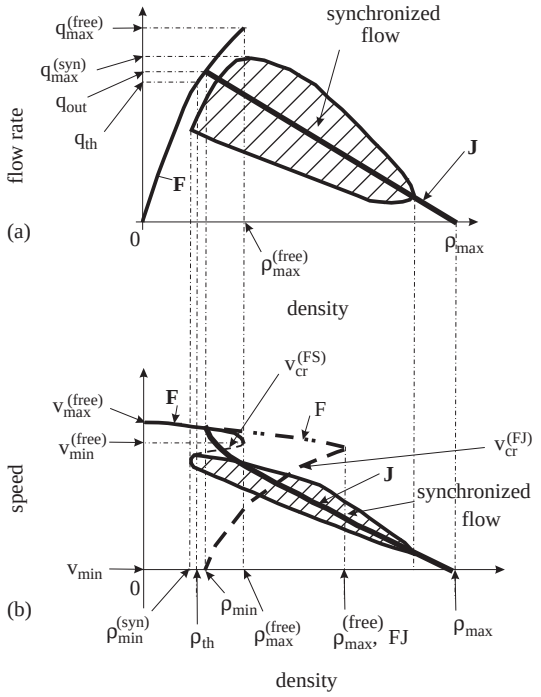

Fig. 12. Explanation of the question why moving jams do not emerge in free flow in the three-phase traffic theory [74]. (a) A qualitative concatenation of states of free flow (curve F), steady states of synchronized flow (hatched region), the critical branch $v_{cr}^{(FS)}$ with Line J in the flow–density plane. (b) states in speed–density plane related to (a). In (a,b) states of free flow (curve F), steady states of synchronized flow (hatched region) and the critical branch $v_{cr}^{(FS)}$ are taken from Fig. 10(a) and (b), respectively.

The density of free flow $\rho^{(free)}$, which is related to these hypothetical free flow states $\rho_{max}^{(free)} < \rho^{(free)} \leqslant \rho_{max,\ FJ}^{(free)}$, cannot be reached in reality (states F', Fig. 12(b)). This is because at the density $\rho^{(free)} = \rho_{max}^{(free)}$ synchronized flow must occur. There should be a critical branch $v_{cr}^{(FJ)}$ in the speed–density plane (Fig. 12(b)). The critical branch $v_{cr}^{(FJ)}$ gives the speed inside the critical perturbation, i.e., it determines the critical amplitude of the critical local perturbation $\Delta v_{cr}^{(FJ)}$ for the $F \to J$ transition: $\Delta v_{cr}^{(FJ)} = v^{(free)} - v_{cr}^{(FJ)}$. Here $v^{(free)}$ is the speed in free flow.

At the critical density $\rho_{max,\ FJ}^{(free)}$ the critical amplitude of the critical perturbation for the $F \to J$ transition is zero: $\Delta v_{cr}^{(FJ)} = 0$. Thus, at the density $\rho^{(free)} = \rho_{max,\ FJ}^{(free)}$ the branch $v_{cr}^{(FJ)}$ should merge with the branch of states of free flow F'. At lower density, $\rho^{(free)} < \rho_{max,\ FJ}^{(free)}$, the critical amplitude of the critical perturbation for the $F \to J$ transition should increase with the decrease in density. At the threshold point of free flow for the $F \to J$ transition, $(\rho_{min},\ q_{out})$, the critical local perturbation with the highest amplitude is needed where the speed is equal to the speed inside a wide moving jam $v_{min} = 0$ (Fig. 12(b)).

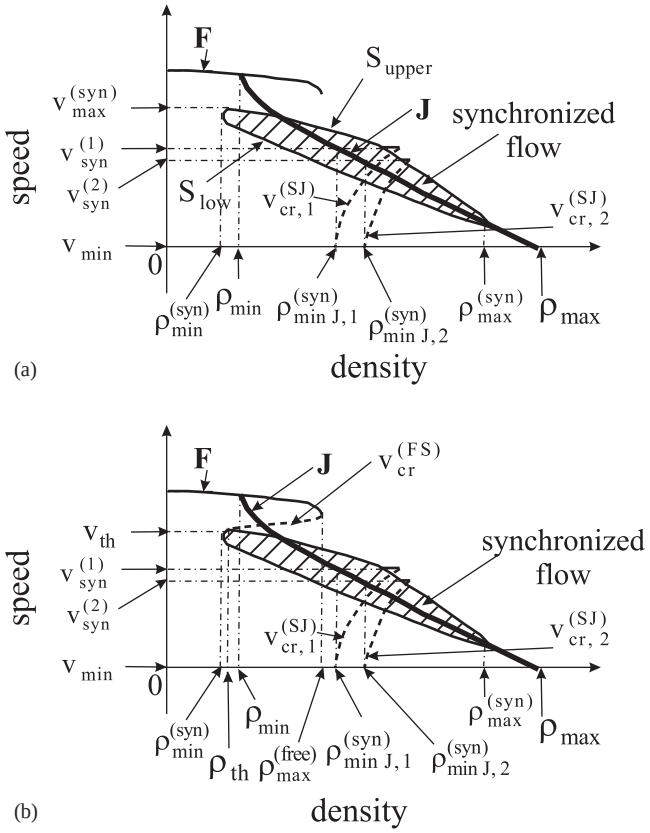

Fig. 13. (a,b) Explanation of the critical branches $v_{cr,1}^{(SJ)}$ and $v_{cr,2}^{(SJ)}$ for the S → J transition for two different steady synchronized flow speeds $v_{syn}^{(1)}$ and $v_{syn}^{(2)}$, respectively. States for free flow (curve F), steady states of synchronized flow (dashed region) and the curve J are taken from Fig. 12(b).

Therefore, we see that the critical amplitude of the local perturbation which is needed for the moving jam emergence in an initial free flow $\Delta v_{cr}^{(FJ)}$ at each density in free flow is higher than the critical amplitude of the local perturbation which is needed for the synchronized emergence in free flow $\Delta v_{cr}^{(FS)}$ (5). This theoretical conclusion may explain why the spontaneous emergence of moving jams is not observed in free flow [74,76].

4.5. Explanation of $F \to S \to J$ transitions: double Z-shaped traffic flow characteristics

Let us first show that the S → J transition considered in Section 4.2.1 is related to a Z-shaped function in the speed–density plane (Fig. 13(a)). To explain this hypothesis, we consider the speeds $v_{syn}^{(1)}$ and $v_{syn}^{(2)}$ from Fig. 7. The threshold densities $\rho_{min\ J,1}^{(syn)}$

and $\rho^{(\mathrm{syn})}_{\mathrm{min}\,J,2}$ for the S → J transition (Fig. 7) correspond to the intersection points between the horizontal line of the speeds $v^{(1)}_{\mathrm{syn}}$ and $v^{(2)}_{\mathrm{syn}}$ with the curve J in the speed–density plane (Fig. 13(a)). Because there is an infinite number of different speeds in synchronized flow, there is also an infinite number of threshold densities for these different synchronized flow speeds. There is also an infinite number of the related critical branches $v^{(\mathrm{SJ})}_{\mathrm{cr}}$ each of then gives the speed inside the critical local perturbation for S → J transition (only two of them, $v^{(\mathrm{SJ})}_{\mathrm{cr},\,1}$ and $v^{(\mathrm{SJ})}_{\mathrm{cr},\,2}$, are shown in Fig. 13(a) for the speeds $v^{(1)}_{\mathrm{syn}}$ and $v^{(2)}_{\mathrm{syn}}$, respectively). The critical amplitude of the critical perturbation in the speed for the S → J transition is equal to $\Delta v^{(\mathrm{SJ})}_{\mathrm{cr}} = v_{\mathrm{syn}} - v^{(\mathrm{SJ})}_{\mathrm{cr}}$, where v_{syn} is one of the infinite possible synchronized flow speeds.

At the threshold density the critical branch $v^{(\mathrm{SJ})}_{\mathrm{cr}}(\rho)$ for the S → J transition ($v^{(\mathrm{SJ})}_{\mathrm{cr}}(\rho)$ gives the speed inside the critical perturbation as a function of the density) should merge with the speed inside a wide moving jam $v_{\mathrm{min}} = 0$ in the speed–density plane (points $(\rho^{(\mathrm{syn})}_{\mathrm{min}\,J,1}, 0)$ and $(\rho^{(\mathrm{syn})}_{\mathrm{min}\,J,2}, 0)$ in Fig. 13(a)). This means that at the threshold density the critical amplitude of the critical perturbation in the speed for the S → J transition is a high value: $\Delta v^{(\mathrm{SJ})}_{\mathrm{cr}} = v_{\mathrm{syn}}$.

At critical points for the S → J transition, where the critical amplitude $\Delta v^{(\mathrm{SJ})}_{\mathrm{cr}} = 0$, the critical branch $v^{(\mathrm{SJ})}_{\mathrm{cr}}(\rho)$ merges with a horizontal line in the speed–density plane. This horizontal line is related to a given synchronized flow speed v_{syn} (e.g., the horizontal lines are related to the speeds $v^{(\mathrm{SJ})}_{\mathrm{cr},\,1}$ and $v^{(\mathrm{SJ})}_{\mathrm{cr},\,2}$ in Fig. 13(a) for the critical branches $v^{(\mathrm{SJ})}_{\mathrm{cr},\,1}$ and $v^{(\mathrm{SJ})}_{\mathrm{cr},\,2}$, respectively). The synchronized flow states (dashed region in Fig. 13(a)), the critical branch $v^{(\mathrm{SJ})}_{\mathrm{cr}}$, and the line $v_{\mathrm{min}} = 0$ for the speed inside a wide moving jam give together a Z-shaped characteristic for the S → J transition (Fig. 13(a)). This Z-characteristic is indeed found in the microscopic three-phase traffic theory [87].

If now we add the critical branch for the F → S transition $v^{(\mathrm{FS})}_{\mathrm{cr}}(\rho)$ which has been considered in Section 4.3.1 to Fig. 13(a), we get a *double Z-shaped* characteristic of the speed on the density (Fig. 13(b)). This explains the empirical result [74] that moving jam spontaneously emerge due to a sequence (cascade) of two phase transitions, first the F → S transition, and later the S → J transition (the F → S → J-transitions). The double Z-shaped characteristic consists of states of free flow F, the critical branch $v^{(\mathrm{FS})}_{\mathrm{cr}}(\rho)$ which gives the speed inside the critical perturbation for the F → S transition, the two-dimensional region of steady states of synchronized flow, the infinite number of critical branches $v^{(\mathrm{SJ})}_{\mathrm{cr}}(\rho)$ which give the speeds inside the critical perturbations for the S → J transition for each of the synchronized flow speeds, and the line $v_{\mathrm{min}} = 0$ which gives the speed inside wide moving jams. The double Z-characteristic is also confirmed in the microscopic three-phase traffic theory [87].

4.6. Phase transitions and patterns at highway bottlenecks

4.6.1. Deterministic perturbation and local breakdown at bottlenecks

Both the phase transitions and spatial–temporal patterns which occur at highway bottlenecks possess important peculiarities in comparison with a homogeneous (without bottlenecks) road considered above. In particular, the following hypothesis is related to this case [79]:

The density of the probability of the $F \to S$ transition for a given time interval, p_{FS} (the probability of the $F \to S$ transition per km) depends on the highway location. The density of this probability has a maximum at the bottleneck (see Fig. 1(b) in Ref. [79]).

This hypothesis is related to the empirical result that the onset of congestion (the $F \to S$-transition) is usually observed at bottlenecks. This can also be explained from driver's behavior. Let us consider a road where only one bottleneck exists. Upstream and downstream of the bottleneck the road has the same characteristics and it is homogeneous. A feature of the bottleneck is that if the flow rate in free flow at the bottleneck is high enough, then in the vicinity of the bottleneck each driver should slow down, i.e., decrease the speed. Thus, a *deterministic* local perturbation in the average speed appears at the bottleneck. This perturbation is motionless and permanent because it is localized at the highway bottleneck: The bottleneck forces all drivers permanently to slow down at approximately the same road location.

The deterministic perturbation is motionless. Thus, the whole flow rate across the road (when taking into account all possible on- and off-ramps) does not depend on the highway location, i.e., this whole flow rate remains the same inside the perturbation at the bottleneck and downstream of the bottleneck. Because inside the deterministic perturbation in free flow the speed is lower than the speed away from the bottleneck but the flow rate does not change, the density must be higher inside the perturbation. This explains why the density of the probability of the $F \to S$ transition (the probability of the $F \to S$ transition per km) should have a maximum at the bottleneck [79]. This hypothesis is also confirmed by results of numerical simulations made in Ref. [87].

Further we consider a bottleneck due to an on-ramp where the flow rate to the on-ramp is q_{on} and the flow rate on the main road downstream of the on-ramp is q_{on}. Consequently, if free flow is at the bottleneck then the flow rate downstream of the bottleneck is equal to

$$q_{sum} = q_{in} + q_{on} \,. \tag{13}$$

The higher the vehicle density in free flow, the higher the amplitude of the local deterministic perturbation. The growth of this deterministic perturbation should have a limit: There should be some critical amplitude of the deterministic perturbation in the speed. When this critical deterministic perturbation at the on-ramp is achieved the *deterministic local speed breakdown* (the $F \to S$ transition) spontaneously occurs: The speed decreases and the density increases avalanche-like at the location of the initial deterministic perturbation. This hypothesis is also confirmed by results of numerical simulations [87]. This deterministic $F \to S$ transition is realized even if no random perturbations (fluctuations) would be in free traffic flow (the deterministic $F \to S$ transition is symbolically marked by dotted arrows $K_{determ} \to M_{determ}$ in Fig. 14(a,b)). This phase transition together with the return $S \to F$ transition (arrow $N \to P$ in Fig. 14(a,b)) may explain the well-known hysteresis in traffic at bottlenecks.

It must be stressed that the deterministic $F \to S$ transition occurs at a *lower* flow rate $q_{sum} = q^{(B)}_{cr,\,FS}$ downstream of the on-ramp than the critical traffic variables on a homogeneous road:

$$q^{(B)}_{cr,\,FS} < q^{(free)}_{max} \,. \tag{14}$$

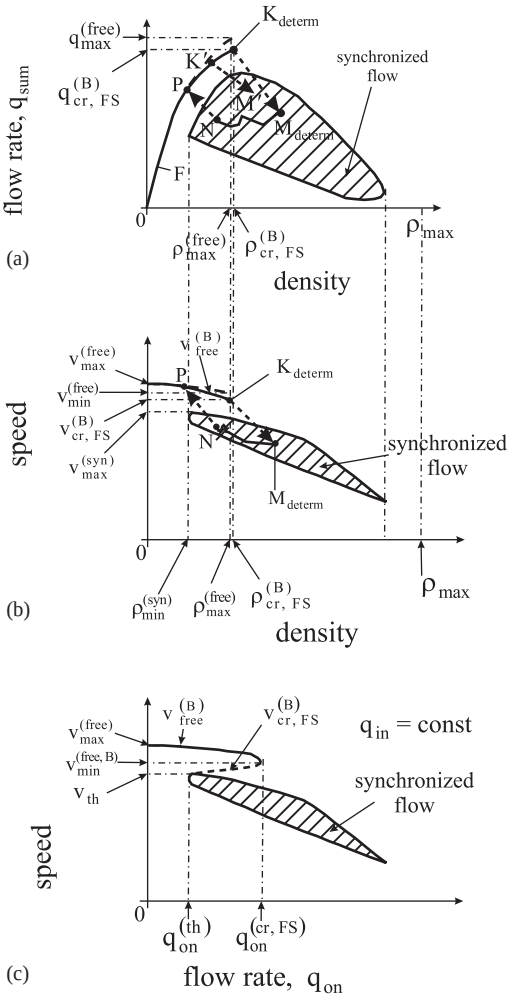

Fig. 14. Qualitative illustration of the deterministic breakdown phenomenon (the deterministic F → S transition) in an initial free flow at a freeway bottleneck: (a) States free flow (solid curves F in (a) and $v_{free}^{(B)}$ in (b) related to the bottleneck; dashed branches show states of free flow on a homogeneous road) and steady states of synchronized flow (dashed region) in the flow–density plane. (b) related to speed–density characteristics of (a). The steady states of synchronized flow are taken from Fig. 1(a). (c) Qualitative illustration of a Z-shaped speed–flow characteristic for the spontaneous breakdown phenomenon (the spontaneous F → S transition) in an initial free flow at a freeway bottleneck due to an on-ramp. States free flow (curve $v_{free}^{(B)}(q_{on})$) and steady states of synchronized flow (dashed region) in the flow–density plane together with the critical branch $v_{cr,FS}^{(B)}(q_{on})$ (dashed curve) which gives the speed inside the critical perturbation.

This formula may explain the empirical result that the breakdown phenomenon (the $F \to S$ transition) is most frequently observed at freeway bottlenecks.

4.6.2. Double Z-characteristics at bottlenecks

Real *random* perturbations in free flow in the bottleneck vicinity can lead to the occurrence of the spontaneous $F \to S$ transition even if the critical point of free flow has not yet been achieved, i.e., at

$$q_{\text{sum}} < q^{(\text{B})}_{\text{cr, FS}} \quad (\rho^{(\text{B})}_{\text{free}}(q_{\text{sum}}) < \rho^{(\text{B})}_{\text{cr, FS}}) \,. \tag{15}$$

This spontaneous $F \to S$ transition is symbolically marked by dotted arrows $K' \to M'$ in Fig. 14(a). Thus, as well as in the case of a hypothetical homogeneous road (Section 4.3.1) we get a Z-shape characteristic of the speed on the density for the onset of congestion, i.e., for the $F \to S$ transition at a bottleneck due to the on-ramp (Fig. 14(b)). In this case, however, the density at the bottleneck is a function of two flow rates q_{on} and q_{in}.

For this reason, it is more convenient to use another Z-characteristic which is related to the dependence of the speed at the bottleneck on the flow rate to the on-ramp q_{on} at a given flow rate q_{in} (Fig. 14(c)). In this Z-characteristic, the branch $v^{(\text{B})}_{\text{free}}(q_{\text{on}})$ gives the speed inside the deterministic perturbation in free flow at the bottleneck as a function of the flow rate q_{on} whereas the branch $v^{(\text{B})}_{\text{cr, FS}}(q_{\text{on}})$ gives the speed inside the critical perturbation with the critical amplitude:

$$\Delta v^{(\text{B})}_{\text{cr, S}} = v^{(\text{B})}_{\text{free}} - v^{(\text{B})}_{\text{cr, FS}} \,. \tag{16}$$

These Z-characteristics at the on-ramp are confirmed in numerical simulations [87].

It is well-known that the breakdown phenomenon at the bottleneck possesses a probabilistic nature [9,11]. This also follows from the Z-characteristic for the $F \to S$ transition (the breakdown phenomenon) at the bottleneck. Indeed, at the critical point $q_{\text{sum}} = q^{(\text{B})}_{\text{cr, FS}}$ (Fig. 14(a)) the critical amplitude of the local perturbation needed for the $F \to S$ transition is zero. Thus, the probability of the $F \to S$ transition at the bottleneck, $P^{(\text{B})}_{\text{FS}}|_{q_{\text{sum}}=q^{(\text{B})}_{\text{cr, FS}}} = 1$. $P^{(\text{B})}_{\text{FS}}$ is defined similar to the probability P_{FS} (8) (see Section 5.2 below). However, rather than a highway section of length L_{ob} which has been considered for the definition of P_{FS} on the homogeneous road, $P^{(\text{B})}_{\text{FS}}$ is the mean probability of the $F \to S$ transition for a given time interval T_{ob} on the main road in the bottleneck vicinity. Because, the critical amplitude of the local perturbation $\Delta v^{(\text{B})}_{\text{cr, FS}}(q_{\text{on}})$ (16) increases when the flow rate q_{on} and consequently the flow rate q_{sum} (13) decrease (Fig. 14(c)), the probability $P^{(\text{B})}_{\text{FS}}$ should decrease. Such dependence of the mean probability of the $F \to S$ transition of the flow rate is observed in empirical observations [11].

After the $F \to S$ transition at a bottleneck has occurred, synchronized flow appears at the bottleneck. The upstream front of this synchronized flow propagates upstream, i.e., the region of synchronized flow is widening. If the speed in the synchronized flow becomes low enough, moving jams can emerge in this synchronized flow. The moving jam emergence occurs away from the bottleneck and it is independent of the reason of the synchronized flow occurrence. This hypothesis is confirmed by empirical results [76]. The $S \to J$ transition in this synchronized flow should also

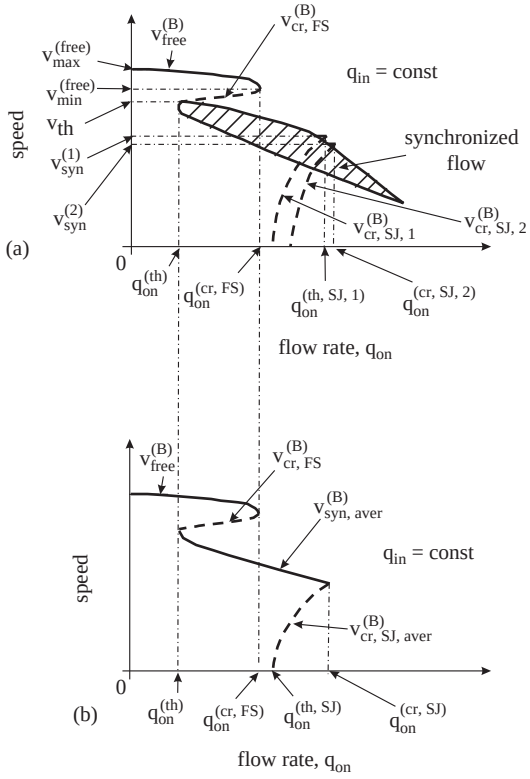

Fig. 15. Qualitative illustration of all possible phase transitions and the double Z-shaped characteristic of traffic flow at the on-ramp. (a) States of free flow $v_{\mathrm{free}}^{(\mathrm{B})}(q_{\mathrm{on}})$, the critical branch $v_{\mathrm{cr,\,FS}}^{(\mathrm{B})}(q_{\mathrm{on}})$ which gives the speed inside the critical local perturbation for the spontaneous $\mathrm{F}\to\mathrm{S}$ transition, a two-dimensional region of steady states of synchronized flow (dashed region), two critical branches $v_{\mathrm{cr,\,SJ},\,i}^{(\mathrm{B})}(q_{\mathrm{on}})$, $i=1,2$ which gives the speed inside the critical local perturbation for the spontaneous $\mathrm{S}\to\mathrm{J}$ transitions for the related two speeds of synchronized flow $v_{\mathrm{syn},\,i}^{(1)}$, $i=1,2$, and the speed $v_{\min}=0$ inside wide moving jams. (b) A simplified double Z-shaped speed–flow characteristic related to (a) where all different synchronized flow speeds at a given flow rate q_{on} are averaged. In (a) states of free flow (the curve $v_{\mathrm{free}}^{(\mathrm{B})}(q_{\mathrm{on}})$), states of synchronized flow (dashed region), and the critical branch $v_{\mathrm{cr,\,FS}}^{(\mathrm{B})}(q_{\mathrm{on}})$ are taken from Fig. 14(c).

possess the same Z-characteristic which has been considered for the homogeneous road in Section 4.5 (Fig. 13(a)). This hypothesis is also confirmed by numerical simulations [87].

This consideration allows us also to expect a relatively complex double Z-shaped traffic flow characteristic at the bottleneck (Fig. 15(a)). On this summarized figure critical branches for the $\mathrm{F}\to\mathrm{S}$ transition and the $\mathrm{S}\to\mathrm{J}$ transition are shown. The double Z-characteristic at the bottleneck consists of

(i) states of free flow $v_{\mathrm{free}}^{(\mathrm{B})}(q_{\mathrm{on}})$,

(ii) the critical branch $v_{\mathrm{cr,\ FS}}^{(\mathrm{B})}(q_{\mathrm{on}})$ which gives the speed inside the critical local perturbation for the spontaneous $\mathrm{F}\to\mathrm{S}$ transition as a function of the flow rate q_{on},

(iii) a 2D region of steady states of synchronized flow (dashed region),

(iv) an infinite number of critical branches $v_{\mathrm{cr,\ SJ}}^{(\mathrm{B})}(q_{\mathrm{on}})$ each of them gives the speed inside the critical local perturbation for the spontaneous $\mathrm{S}\to\mathrm{J}$ transition as a function of the flow rate q_{on} for a given synchronized flow speed (two of these critical branches $v_{\mathrm{cr,\ SJ,\ }i}^{(\mathrm{B})}(q_{\mathrm{on}})$, $i=1,2$ for the related two speeds of synchronized flow $v_{\mathrm{syn}}^{(i)}$, $i=1,2$ are shown in Fig. 15(a)) and

(v) the speed $v_{\min}=0$ inside wide moving jams.

If we simplify Fig. 15(a) by an averaging of all different states of synchronized flow at a given flow rate q_{on} to the one speed we can find much more simple double Z-shaped traffic flow characteristic at the bottleneck (Fig. 15(b)).

The hypothesis about the double Z-characteristic at the bottleneck explains the empirical $\mathrm{F}\to\mathrm{S}\to\mathrm{J}$ transitions [76]. This hypothesis is also confirmed by numerical simulations in the frame of the three-phase traffic theory [87]. In contrast, *no* traffic flow theories and models in the fundamental diagram approach [31–34] can show the empirical $\mathrm{F}\to\mathrm{S}\to\mathrm{J}$ transitions and the theoretical double Z-characteristics for traffic flow.

4.6.3. Diagram of congested patterns

The qualitative difference of traffic flow theories and models in the fundamental diagram approach [31–34] and of the three-phase traffic theory [12,74,78,80,84] can essentially clearly be seen if the diagrams of congested patterns at highway bottlenecks which should occur in these different approaches are compared [76,85–88].

Recall that in the diagram of congested patterns at the on-ramp in the fundamental diagram approach which has first been found by Helbing et al. in Ref. [62] diverse congested patterns are possible depending on the initial flow rate q_{in} on a highway upstream of the bottleneck and on the flow rate to the on-ramp q_{on}.

The diagram of the congested patterns at on-ramps within the three-phase traffic theory has first been postulated by the author [76,103] based on qualitative considerations and then derived by Kerner and Klenov based on their microscopic traffic flow model [85]. It has already been shown in [85,86] that this diagram is totally qualitatively different from the diagram of congested patterns at on-ramps in the fundamental diagram approach [62,63,66,32–34].

In particular, in contrast to the fundamental diagram approach [32,33], in the diagram of congested patterns within the three-phase traffic theory [76,85,103,114] at a high enough flow rate to the on-ramp moving jams always spontaneously emerge in synchronized flow upstream of the bottleneck rather than HCT. At a low enough flow rate to the on-ramp rather than moving jams synchronized flow of higher vehicle speed can occur without an occurrence of moving jams. These theoretical results are in agreement with the related results of empirical observations [76] (see the empirical scheme 2 and the related discussion in Section 3).

A qualitative derivation of the diagram of congested patterns at on-ramps has already been made in Refs. [76,103]. Here a brief consideration of this diagram which is necessary for further consideration will be made.

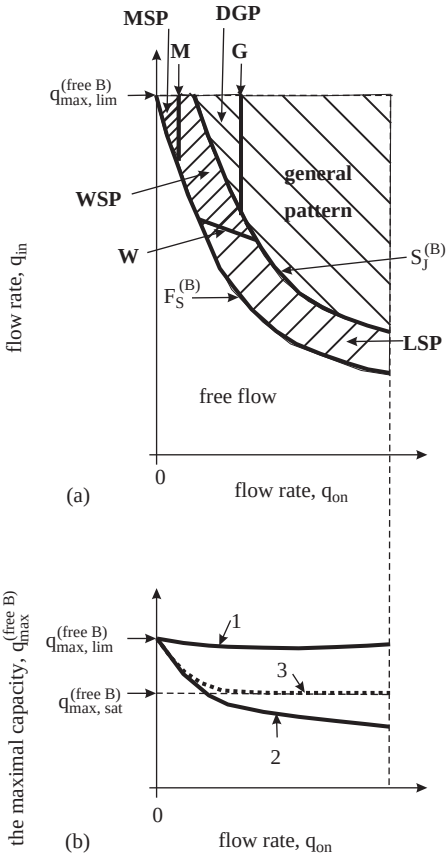

Fig. 16. Diagram of congested patterns at highway bottlenecks on a multi-lane highway in the three-phase traffic theory (a) [76,85,103,114] and possible dependencies of the maximum highway capacity in free flow at the effective location of a bottleneck $q_{max}^{(free\ B)}$ on the flow rate to the on-ramp q_{on} (b).

There are two main boundaries in the diagram of congested patterns at highway bottlenecks: $F_S^{(B)}$ and $S_J^{(B)}$ [76,85,103]. Below and left of the boundary $F_S^{(B)}$ free flow is realized (Fig. 16(a)). Between the boundaries $F_S^{(B)}$ and $S_J^{(B)}$ different SPs occur. Right of the boundary $S_J^{(B)}$ wide moving jams spontaneously emerge in synchronized flow upstream of the bottleneck, i.e., a GP occur [76,85,103].

To explain the boundary $F_S^{(B)}$, let us first consider the following hypothesis: Let us assume that the flow rate q_{in} is high enough but the flow rate q_{on} is extremely small, i.e.,

$$q_{on} \to 0 \quad \text{but} \quad q_{on} \neq 0\,. \tag{17}$$

Under condition (17) *the critical flow rate* $q_{max}^{(free\ B)}|_{q_{on}\to 0} = q_{max,lim}^{(free\ B)}$ *for the* $F \to S$ *transition at the bottleneck is lower than the critical flow rate* $q_{max}^{(free)}$ *for the* $F \to S$

transition on the homogeneous road:

$$q_{\max,\lim}^{(\text{free B})} < q_{\max}^{(\text{free})} . \tag{18}$$

Under condition (17), the influence of q_{on} on the flow rate q_{sum} (13) can be neglected: $q_{\text{sum}} = q_{\text{in}}$. In the case (17), time intervals between single vehicles squeezing onto the main road from the on-ramp can be large enough. During these time intervals the main road can be considered as the homogeneous one, because it is non-disturbed by the on-ramp. However, when single vehicles squeeze onto the main road, they cause time-limited *additional* random disturbances in free flow on the main road in the on-ramp vicinity. These additional random perturbations can obviously cause the F → S transition at the on-ramp at a lower flow rate q_{in} than would be the case at $q_{\text{on}} = 0$, i.e., on the homogeneous road. This explains the hypothesis (18) which is confirmed by numerical simulations [87].

The boundary $F_{\text{S}}^{(\text{B})}$ can be found from the following qualitative consideration. The critical amplitude of the critical local perturbation for the F → S transition at the bottleneck tends to zero if the flow rate approaches a critical point of free flow. In the case of the bottleneck due to the on-ramp, there are an infinite number of these critical points of free flow. These critical points are related to an infinite number of the critical flow rates downstream of the bottleneck $q_{\text{sum}}(q_{\text{on}}, q_{\text{in}})|_{F_{\text{S}}^{(\text{B})}}$ which correspond to the boundary $F_{\text{S}}^{(\text{B})}$ in the diagram of congested patterns. At $q_{\text{on}} \to 0$ (17) the critical flow rate $q_{\text{sum}}|_{F_{\text{S}}^{(\text{B})}}$ in the related critical point of free flow is equal to $q_{\text{in}} = q_{\max,\lim}^{(\text{free B})}$. Thus, the F → S transition at the on-ramp must occur in this critical point. The higher q_{on}, the higher the amplitude of the deterministic local perturbation caused by the on-ramp. Therefore, the higher q_{on}, the lower the flow rate in free flow on the main road upstream of the bottleneck q_{in} should be (in comparison with $q_{\text{in}} = q_{\max,\lim}^{(\text{free B})}$) at which the F → S-transition at the on-ramp must occur. This explains the form of the boundary $F_{\text{S}}^{(\text{B})}$ for the spontaneous F → S transition at the bottleneck in Fig. 16(a).

The boundary $S_{\text{J}}^{(\text{B})}$ is determined by the wide moving jam emergence in synchronized flow (i.e., the S → J transition) upstream of the on-ramp. On the one hand, between the boundaries $F_{\text{S}}^{(\text{B})}$ and $S_{\text{J}}^{(\text{B})}$ the vehicle speed in a SP should decrease when q_{on} increases. On the other hand, the lower the vehicle speed in synchronized flow, the more probability of the wide moving jam emergence in this synchronized flow. Thus, in comparison with $F_{\text{S}}^{(\text{B})}$, the boundary $S_{\text{J}}^{(\text{B})}$ should be shifted to the right in the flow–flow plane (Fig. 16(a)).

Between the boundaries $F_{\text{S}}^{(\text{B})}$ and $S_{\text{J}}^{(\text{B})}$, the higher q_{in} is, the higher the probability that the flow rate in synchronized flow in the SP is lower than q_{in} and the length of the SP is continuously increasing over time: At higher q_{in} a widening SP (WSP) and at lower q_{in} a localized SP (LSP) occurs (Fig. 16(a)). The flow rate inside a WSP is lower than q_{in}. Therefore, the upstream WSP front (boundary), which separates free flow upstream and synchronized flow downstream, is continuously widening upstream. The mean flow rate inside the LSP is equal to q_{in}. For this reason, the upstream LSP front is not continuously widening upstream: The width of the LSP is spatially limited. However, this LSP width can show oscillations over time [85,86].

Right of the boundary $F_S^{(B)}$ and left of the line M one or a sequence of moving SPs (MSP) emerge upstream of the on-ramp (the region marked "MSP" in Fig. 16(a)). In contrast to a wide moving jam, inside a MSP both the vehicle speed (40–70 km/h) and the flow rate are high. Besides, the velocity of the downstream front of a MSP is *not* a characteristic parameter. This velocity can change in a wide range in the process of the MSP propagation or for different MSPs. In some cases it has been found that after the MSP is far away from the on-ramp, the pinch effect (the self-compression of synchronized flow) occurs inside the MSP and a wide moving jam can be formed there [85].

Right of the boundary $S_J^{(B)}$ and left of the line *G the dissolving general pattern* or DGP for short occurs (the region marked "DGP" in Fig. 16(a)). In the DGP, after a wide moving jam in synchronized flow of the congested pattern has been formed, the forming GP dissolves over time. As a result of this GP dissolving process, the GP transforms into one of the SP, or free flow occurs at the bottleneck [85,86].

The following empirical results confirm the diagram of congested patterns in the three-phase traffic flow theory (Fig. 16(a)):

(i) All types of SPs and GPs are found in empirical observations (see Ref. [76] and Section 2.2).

(ii) In the diagram (Fig. 16(a)), GP exists in the most part of the flow rates q_{on} and q_{in} where congestion occurs. This is related to the empirical result that GP is the most frequent type of congested patterns at isolated highway bottlenecks [76].

(iii) Corresponding to the diagram, in empirical investigations GP transforms into a SP when the flow rate q_{on} decreases [76].

(iv) Corresponding to the diagram, in empirical investigations GP does not transform into another type of congested pattern when the flow rate q_{on} increases [76].

The hypothesis about the diagram of congested patterns [76] is also confirmed by the diagrams found in the microscopic three-phase traffic flow theories [85–87].

In contrast to this, it has been shown in the empirical study [76] that no sequences of the congested pattern transformation and no theoretical congested states which have been predicted in the pattern diagram in the fundamental diagram approach [62,66,32,33] have been observed at isolated bottlenecks.

5. Probabilistic theory of highway capacity

The determination of highway capacity is one of the most important applications of any traffic theory. Empirical observations show that the speed breakdown at a bottleneck (the breakdown phenomenon) is in general accompanied by a drop in highway capacity (see e.g., Refs. [8,7]). Here we give a qualitative theory of highway capacity and of the capacity drop which follows from the three-phase traffic theory.

However, firstly recall, how the breakdown phenomenon looks like in the fundamental diagram approach. From a numerical analysis of a macroscopic traffic flow model within the fundamental diagram approach Kerner and Konhäuser found in 1994 [46]

that free flow is metastable with respect to the formation of wide moving jams (F → J transition), if the flow rate is equal to or higher than the outflow from a jam, q_{out}. The critical amplitude of a local perturbation in an initial homogeneous free flow, which is needed for the F → J transition, decreases with increasing density: It is maximum at the threshold density $\rho = \rho_{\min}$, below which free flow is stable. The critical amplitude becomes zero at some critical density $\rho = \rho_{\text{cr}} > \rho_{\min}$, above which free flow is linearly unstable. Obviously the higher the amplitude of a random local perturbation the less frequent it is. Hence, the likelihood that the F → J transition occurs in a given time interval should increase with density (or flow rate). The probability should tend to one at the critical density ρ_{cr}.

In 1997 Mahnke et al. [107,57] developed a master equation approach for calculating the probability of the F → J transition on a homogeneous road (i.e., without bottleneck). Based on this approach Kühne et al. [95] confirmed that the probability of the F → J transition in the metastable region is increasing with the flow rate in free flow. They applied this result to explain the breakdown phenomenon at a highway bottleneck. For a recent comprehensive discussion of the breakdown phenomenon in CA-models and in the Krauß et al. model in the fundamental diagram approach see also [96]. The theories in Refs. [46,32,107,57,95,96] belong to the fundamental diagram approach.

In contrast to these results, in the three-phase traffic theory [12,80,84] it is postulated that metastable states of free flow decay into synchronized flow (F → S transition) rather than wide moving jams (F → J transition). In particular, even the upper limit of free flow ($q_{\text{max,lim}}^{(\text{free B})}$ in Fig. 16) is related to the F → S transition: In this limit point the probability of the F → S transition should be equal to one whereas the probability of the emergence of a moving jam (F → J transition) should be very small (Fig. 9(c), curve F_{J}). Thus, in this theory the breakdown phenomenon in free traffic is related to the F → S transition rather than to an emergence of moving jams.

Highway capacity depends on whether a homogeneous road (without bottlenecks) or a highway bottleneck is considered.

5.1. Homogeneous road

On a homogeneous (without bottlenecks) multi-lane road, highway capacity depends on which traffic phase the traffic is in [74]: (i) The maximum highway capacity in the traffic phase "free flow" is equal to the maximum possible flow rate in free flow, $q_{\max}^{(\text{free})}$. (ii) The maximum highway capacity in the traffic phase "synchronized flow" is equal to the maximum possible flow rate in synchronized flow, $q_{\max}^{(\text{syn})}$. (iii) The maximum highway capacity downstream of the traffic phase "wide moving jam" is equal to the flow rate in the wide moving jam outflow, q_{out}. Because of the first order phase transitions between the traffic phases each of these maximum highway capacities has a probabilistic nature.

In particular, the probabilistic nature of the highway capacity in the traffic phase "free flow" means the following [12]:

(1) At the flow rate $q = q_{\max}^{(\text{free})}$ the probability of the spontaneous F → S transition for a given time interval T_{ob} and for a given highway section length L_{ob}, $P_{\text{FS}}\big|_{q=q_{\max}^{(\text{free})}} = 1$

(7). In other words, the maximum capacity $q = q_{\max}^{(\text{free})}$ depends on T_{ob} (at least in some range of T_{ob}).

(2) There is the threshold flow rate q_{th} (the threshold density ρ_{th}) for the $\text{F} \to \text{S}$ transition. Below the threshold point $(\rho_{\text{th}}, q_{\text{th}})$, i.e., at the flow rate $q < q_{\text{th}}$ (Fig. 9(a)) the probability of the $\text{F} \to \text{S}$ transition $P_{\text{FS}}|_{q<q_{\text{th}}} = 0$ (9).

(3) If the flow rate in free flow q is within the range $[q_{\text{th}}, q_{\max}^{(\text{free})}]$ then the higher the flow rate q, the higher the probability the $\text{F} \to \text{S}$ transition P_{FS} is. Thus, the attribute of this *probabilistic highway capacity* is the probability $1 - P_{\text{FS}}$ that free flow remains on a road section of the length L_{ob} during the time interval T_{ob} of the observation of this capacity in free flow.

The hypotheses of the three-phase traffic theory are confirmed by empirical findings of characteristics of the wide moving jam outflow (e.g., Refs. [77,113]) and by numerical results of a microscopic three-phase traffic flow theory [85–87].

However, if a bottleneck exist on the road, then a much more complicated non-linear phenomena determine highway capacity.

5.2. Highway capacity in free flow at bottleneck

In the three-phase traffic theory, the breakdown phenomenon at a highway bottleneck is explained by the $\text{F} \to \text{S}$ transition at the bottleneck [79]. Due to the bottleneck the road is spatially non-homogeneous, i.e., highway capacity can depend on a highway location [14,15].

The $\text{F} \to \text{S}$ transition occurs at the bottleneck during a given time interval T_{ob} if the flow rate q_{in} and the flow rate q_{on} are related to the boundary $F_{\text{S}}^{(\text{B})}$ in the diagram of congested patterns (Fig. 16(a)). Thus, there is *an infinite multitude* of maximum freeway capacities of free flow at the bottleneck which are given by the points on the boundary $F_{\text{S}}^{(\text{B})}$. We designate these capacities $q_{\max}^{(\text{free B})}$:

$$q_{\max}^{(\text{free B})} = q_{\text{sum}}|_{F_{\text{S}}^{(\text{B})}} , \tag{19}$$

where q_{sum} (13) is the flow rate in free flow downstream of the bottleneck.

The capacities $q_{\max}^{(\text{free B})}$ (19) depend on the flow rate on the main road upstream of the on-ramp q_{in} and the flow rate to the on-ramp q_{on}, exactly on the values q_{in} and q_{on} at the boundary $F_{\text{S}}^{(\text{B})}$ in the diagram of congested patterns at the on-ramp (Fig. 16(a)). The capacities (19) are the maximum capacities related to a given time interval T_{ob}. These freeway capacities as well as the boundary $F_{\text{S}}^{(\text{B})}$ are functions of the time of the observation of the $\text{F} \to \text{S}$ transition, T_{ob}.

To find the maximum capacities $q_{\max}^{(\text{free B})}$ (19), the probability of the $\text{F} \to \text{S}$ transition at the bottleneck $P_{\text{FS}}^{(\text{B})}$ for a given time interval T_{ob} should be studied.

This probability $P_{\text{FS}}^{(\text{B})}$ is defined as follows. A large number of different realizations, N_{FS}, are performed where the $\text{F} \to \text{S}$ transition in an initial free flow at the bottleneck is studied. Each of the realizations should be made at the same flow rates q_{on} and q_{in}, other initial conditions, and during the same time interval T_{ob} of the observation of the spontaneous $\text{F} \to \text{S}$ transition at the bottleneck. If in n_{FS} of these N_{FS} realizations the

F → S transition occurs, then [7]

$$P_{\mathrm{FS}}^{(\mathrm{B})} = \frac{n_{\mathrm{FS}}}{N_{\mathrm{FS}}} \ . \tag{21}$$

Here the remark in footnote 5 is also valid.

The maximum freeway capacities of free flow at the bottleneck, $q_{\max}^{(\text{free B})}$, are found from the condition that the probability $P_{\mathrm{FS}}^{(\mathrm{B})}$ reaches one during the interval of the observation T_{ob}: [8]

$$P_{\mathrm{FS}}^{(\mathrm{B})}|_{q_{\mathrm{sum}}=q_{\max}^{(\text{free B})}} = 1 \ . \tag{22}$$

Thus, condition (22) determines the critical flow rates $q_{\max}^{(\text{free B})}$ at the boundary $F_{\mathrm{S}}^{(\mathrm{B})}$ in the diagram of congested patterns.

For traffic demand (values q_{on} and q_{in}) which is related to points in some vicinity *below and left of* the boundary $F_{\mathrm{S}}^{(\mathrm{B})}$ in the diagram of congested patterns, i.e., in the free flow region of the diagram the F → S transition nevertheless occurs at the bottleneck during the time interval T_{ob} with a probability $P_{\mathrm{FS}}^{(\mathrm{B})} < 1$. In this case, the more the distant a point (q_{on}, q_{in}) in the diagram of congested patterns is from the boundary $F_{\mathrm{S}}^{(\mathrm{B})}$ (Fig. 16(a)), the lower the probability $P_{\mathrm{FS}}^{(\mathrm{B})}$ is.

The region in the diagram of congested patterns where this probabilistic effect occurs is restricted by a threshold boundary $F_{\mathrm{th}}^{(\mathrm{B})}$ which is below and left of the boundary $F_{\mathrm{S}}^{(\mathrm{B})}$ (Fig. 17(b)): [9] This threshold boundary is related to the *infinite multitude* of threshold flow rates q_{sum} which we designate $q_{\mathrm{th}}^{(\mathrm{B})}(q_{\mathrm{on}}, q_{\mathrm{in}})$. At the threshold boundary $F_{\mathrm{th}}^{(\mathrm{B})}$ the critical amplitudes of the local perturbation for the F → S transition reach maximum values. The limit point on the threshold boundary $F_{\mathrm{th}}^{(\mathrm{B})}$ at $q_{\mathrm{on}} = 0$, i.e., when $q_{\mathrm{sum}} = q_{\mathrm{in}}$ is the threshold point for a homogeneous road: $q_{\mathrm{th}}^{(\mathrm{B})}(0, q_{\mathrm{in}}) = q_{\mathrm{th}}$ where the threshold flow rate for the homogeneous road, q_{th}, has been defined in Section 4.3.1.

Other threshold points on the threshold boundary $F_{\mathrm{th}}^{(\mathrm{B})}$ for $q_{\mathrm{on}} > 0$ are related to the flow rate $q_{\mathrm{in}} < q_{\mathrm{th}}$. These threshold points are defined as follows. Let us consider

---

[7] In empirical observations, the flow rate is averaged over a time-interval T_{av}. In free flow, this averaged flow rate cannot be considered as a time-independent function for usual time intervals $T_{\mathrm{av}} \approx 0.5$–$10$ min. For this reason, another definition of the probability $P_{\mathrm{FS}}^{(\mathrm{B})}$ can be used [11]: Observed flows are divided into groups in some small multiples. The frequency of flows, N_i, in each group I in free flow is studied before the F → S transition occurs at the bottleneck. For each of flow group, I, over a large number of days, the number of instances, n_i is found, in which the flow rate q_i of the F → S transition occurrence at the bottleneck falls within the range of flow group, I. Thus, the probability $P_{\mathrm{FSi}}^{(\mathrm{B})}$ of the F → S transition for the flow group I is [11]:

$$P_{\mathrm{FSi}}^{(\mathrm{B})} = \frac{n_i}{N_i} \ . \tag{20}$$

If the flow rate within each of the intervals T_{av} of the flow rate averaging can be considered as time-independent and traffic control parameters (weather, etc.) on different days of the observations do not change, then the empirical (20) and the theoretical definitions (21) for the bottleneck due to the on-ramp are equivalent at $q_{\mathrm{sum}} = q_i$, $T_{\mathrm{ob}} = T_{\mathrm{av}}$, and the same empirical and theoretical flow rates q_{on}.

[8] Whereas condition (22) determines the theoretical maximum freeway capacities of free flow at the bottleneck which depend on T_{ob}, the condition $P_{\mathrm{FSi}}^{(\mathrm{B})}|_{q_i=q_{\max}^{(\text{free B})}} = 1$ determines empirical maximum freeway capacities of free flow at the bottleneck which depend on T_{av} (see footnote 7).

[9] A more detail consideration of stable and metastable free flow states and of regions of metastability of different congested patterns in the diagram of the patterns has recently been made in Ref. [87].

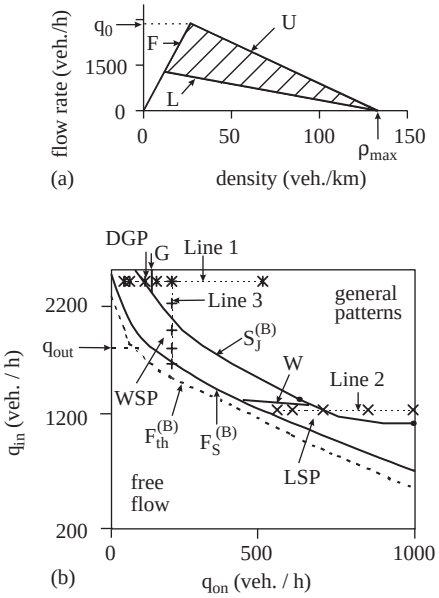

Fig. 17. Steady states in the flow–density plane (a) [86] and the diagram of congested patterns at the on-ramp (b) for the KKW CA-model within the three-phase traffic theory. The model parameters are related to the KKW-1 CA model (parameter set I) in Ref. [86].

a point (q_{on}, q_{in}) in the diagram of congested patterns which is below and left of the boundary $F_{FS}^{(B)}$, however, the $F \to S$ transition occurs in this free flow with the probability $0 < P_{FS}^{(B)} < 1$. If now the flow rate q_{sum} decreases (due to a gradual decrease in the flow rate q_{on} and/or the flow rate q_{in}) the distance of the point (q_{on}, q_{in}) from the critical boundary $F_{FS}^{(B)}$ increases and the mean probability of the $F \to S$ transition $P_{FS}^{(B)}$ decreases. The threshold points $q_{th}^{(B)}(q_{on}, q_{in})$ on the threshold boundary $F_{th}^{(B)}$ are the *minimum* flow rates q_{sum} where the $F \to S$ transition during a given time T_{ob} can *still* occur: If $q_{sum} < q_{th}^{(B)}$, then the $F \to S$ transition cannot occur in free flow at the bottleneck. This means that below and left of the threshold boundary $Fw_{th}^{(B)}$, i.e., when $q_{sum} < q_{th}^{(B)}$ the probability for the $F \to S$ transition $P_{FS}^{(B)}$ is given by the formula:

$$P_{FS}^{(B)}|_{q_{sum} < q_{th}^{(B)}} = 0 \,. \tag{23}$$

This formula is analogous to the formula (9). [10]

[10] In numerical simulations (Section 5.4), the threshold boundary $F_{th}^{(B)}$ can be estimated on another way: The probability $P_{FS}^{(B)}$ is calculated as it has above been defined. However we start calculations of $P_{FS}^{(B)}$ beginning from low enough values $q_{sum}(q_{on}, q_{in})$ when in all N_{FS} realizations no $F \to S$ transitions occur at the on-ramp, i.e., $P_{FS}^{(B)} = 0$. Then, the flow rate q_{sum} increases. The threshold flow rates $q_{th}^{(B)}$ at the threshold boundary $F_{th}^{(B)}$ are approximately related to the *maximum* flow rates q_{sum} at which the condition

$$P_{FS}^{(B)}|_{q_{sum} = q_{th}^{(B)}} = 0 \tag{24}$$

is *still* satisfied.

There can be a number of different dependencies of the maximum freeway capacity in free flow at the bottleneck $q_{\max}^{(\text{free B})}$ on q_{on} (e.g., curves 1, 2, 3 in Fig. 16(b)). In particular, it can be expected that the higher the flow rate to the on-ramp q_{on}, the lower the maximum freeway capacity in free flow at the bottleneck $q_{\max}^{(\text{free B})}$. In this case, the maximum freeway capacity $q_{\max}^{(\text{free B})}$ is a decreasing function of q_{on} (curve 2 in Fig. 16(b)). A decrease of the maximum freeway capacity $q_{\max}^{(\text{free B})}$ at the on-ramp when the flow rate q_{on} increases can have saturation at a high enough flow rate to the on-ramp q_{on} (Fig. 16(b), dotted curve 3). In this case, the maximum freeway capacity $q_{\max}^{(\text{free B})}$ does not reduce below some saturation value $q_{\max,\,\text{sat}}^{(\text{free B})}$, even at a very high q_{on}.

The highest is the capacity of free flow at $q_{\text{on}} = 0$. This case is obviously related to the maximum capacity on a homogeneous (without bottlenecks) road (Fig. 9(a)):

$$q_{\max}^{(\text{free B})}|_{q_{\text{on}}=0} = q_{\max}^{(\text{free})} . \tag{25}$$

However, at a very small flow rate q_{on}, exactly at the limit case (17) we have

$$q_{\max}^{(\text{free B})}|_{q_{\text{on}}\to 0} = q_{\max,\text{lim}}^{(\text{free B})} < q_{\max}^{(\text{free})} \quad \text{at } q_{\text{on}} \neq 0 . \tag{26}$$

Condition (26) is linked to the effect of random perturbations which occur at the on-ramp due to the single vehicles squeezing onto the main road from the on-ramp as has already been explained in Section 4.6.3. For this reason in Fig. 16(b) rather than the flow rate $q_{\max}^{(\text{free})}$ the flow rate $q_{\max,\text{lim}}^{(\text{free B})}$ is shown as the maximum freeway capacity at the bottleneck at the limit case $q_{\text{on}} \to 0$.

Corresponding to (22) and (23), if the flow rate q_{sum} (26) at the effective location of the bottleneck is within the range $[q_{\text{th}}^{(\text{B})}, q_{\max,\text{lim}}^{(\text{free B})}]$ then at a given q_{on} the higher q_{sum}, the higher the mean probability of the F → S transition $P_{\text{FS}}^{(\text{B})}$.

These hypotheses of the three-phase traffic theory are confirmed by empirical findings where the probabilistic nature of freeway capacity has been studied [11] and by numerical results [86,87].

5.3. Highway capacity in congested traffic at bottleneck. Capacity drop

In order to study the highway capacity downstream of the congested bottleneck one has to consider the outflow from a congested bottleneck $q_{\text{out}}^{(\text{bottle})}$ (the discharge flow rate), which is measured downstream of the bottleneck, where free flow conditions are reached.

In the three-phase traffic theory, the discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ is not just a characteristic property of the type of bottleneck under consideration only. It also depends on the type of congested pattern which actually is formed upstream of the bottleneck [79]. Thus, in the three-phase traffic theory, highway capacity in free flow downstream of the congested bottleneck depends on the type of congested pattern upstream of the bottleneck, the pattern characteristics and on parameters of the bottleneck. We call this highway capacity *a congested pattern capacity*.

In the case of an on-ramp, $q_{\text{out}}^{(\text{bottle})}$ is expected to vary with $(q_{\text{on}}, q_{\text{in}})$. Obviously, $q_{\text{out}}^{(\text{bottle})}$ only limits the freeway capacity, if it is smaller than the traffic demand upstream

of the on-ramp, $q_{\rm sum} = q_{\rm in} + q_{\rm on}$, i.e., if the condition

$$q_{\rm out}^{\rm (bottle)}(q_{\rm on}, q_{\rm in}) < q_{\rm sum} \tag{27}$$

is fulfilled. Note that in (27) in contrast to the discharge flow rate $q_{\rm out}^{\rm (bottle)}$, the flow rate $q_{\rm sum}$ is related to free flow conditions at the bottleneck, i.e., when no congested pattern exists upstream of the bottleneck. Under the condition (27), the congested pattern upstream from the on-ramp simply expands, while the throughput remains limited by $q_{\rm out}^{\rm (bottle)}$. For example, if the general pattern (GP) is formed at the bottleneck, an increase of $q_{\rm in}$ does not influence the discharge flow rate $q_{\rm out}^{\rm (bottle)}$. Instead, the width of the wide moving jam, which is mostly upstream in the GP, simply grows.

Thus, in case (27) the congested pattern capacity $q_{\rm cong}^{\rm (B)}$ is equal to $q_{\rm out}^{\rm (bottle)}$:

$$q_{\rm cong}^{\rm (B)} = q_{\rm out}^{\rm (bottle)} . \tag{28}$$

It must be noted that $q_{\rm out}^{\rm (bottle)}$ can strongly depend on the congested pattern type and congested pattern parameters. The congested pattern type and the pattern parameters depend on initial conditions and the flow rates $q_{\rm on}$ and $q_{\rm in}$ (Fig. 16(a)). Thus, the congested pattern capacity $q_{\rm cong}^{\rm (B)}$ implicitly depends on the flow rates $q_{\rm on}$ and $q_{\rm in}$.

The capacity drop is the difference between freeway capacity in free flow at a bottleneck and in a situation, where there is synchronized flow upstream and free flow downstream of the bottleneck (e.g., Ref. [8]).

Assuming that (27) is fulfilled, the capacity drop can be given by

$$\delta q = q_{\rm max}^{\rm (free)} - q_{\rm cong}^{\rm (B)} , \tag{29}$$

where the congested pattern capacity $q_{\rm cong}^{\rm (B)}$ is given by formula (28); $q_{\rm max}^{\rm (free)}$ is the maximum freeway capacity in free flow at $q_{\rm on} = 0$.

If one considers all kinds of congested patterns upstream from a bottleneck then there should be the minimum discharge flow rate which satisfies condition (27). This minimum discharge flow rate should be the characteristic quantity for the type of bottleneck under consideration. We denote this quantity by $q_{\rm min}^{\rm (bottle)}$. The maximum of $q_{\rm out}^{\rm (bottle)}$ (denoted by $q_{\rm max}^{\rm (bottle)}$) is predicted to be the maximum flow rate, which can be realized in synchronized flow, $q_{\rm max}^{\rm (bottle)} = q_{\rm max}^{\rm (syn)}$. To explain the latter condition, recall that the downstream front of a congested pattern at the bottleneck due to the on-ramp separates free flow downstream of the front and synchronized flow upstream of the front. This downstream front is fixed at the bottleneck. Thus, within the front the total flow rate across the road (together with the on-ramp) does not depend on the co-ordinates along the road. Just downstream of the front, i.e., in free flow this flow rate is equal to $q_{\rm max}^{\rm (bottle)}$. Just upstream of the front it is suggested that synchronized flow occurs both on the main road and the on-ramp. The maximum possible total flow rate in synchronized flow is equal to $q_{\rm max}^{\rm (syn)}$ (Fig. 1(a)). Hence, the capacity drop at a bottleneck cannot be smaller than

$$\delta q_{\rm min} = q_{\rm max}^{\rm (free)} - q_{\rm max}^{\rm (syn)} . \tag{30}$$

Note that there may also be another definition of the capacity drop:

$$\delta q = q_{\max}^{(\text{free B})} - q_{\text{cong}}^{(\text{B})} , \tag{31}$$

where the congested pattern capacity $q_{\text{cong}}^{(\text{B})}$ is given by formula (28) and $q_{\max}^{(\text{free B})}$ is given by (19). However, there could be a difficulty in the application of the definition (31): There is an infinite multitude of different maximum freeway capacities in free flow at a bottleneck, $q_{\max}^{(\text{free B})}$ (see formula (19)).

There may be one exception of condition (27): If a LSP occurs both on the main road and on the on-ramp upstream of the merge region of the on-ramp then the discharge flow rate is equal to traffic demand:

$$q_{\text{out}}^{(\text{bottle})}(q_{\text{on}}, q_{\text{in}}) = q_{\text{sum}} . \tag{32}$$

The congested pattern capacity which is related to this LSP should be determined by the maximum discharge flow rate at which the LSP still exists upstream of the on-ramp.

These hypotheses of the three-phase traffic theory are confirmed by empirical findings where the discharge flow rate from spatial–temporal congested patterns at bottlenecks has been studied [76] and by numerical results derived within the three-phase traffic theory which will be presented below.

5.4. Numerical study of congested pattern capacity at on-ramps

Here we confirm and illustrate the general theory of congested pattern capacity presented above based on a numerical simulation of a one-lane KKW cellular automata microscopic traffic flow model within the three-phase traffic theory which has recently been proposed by Kerner et al. [86].

5.4.1. KKW cellular automata traffic flow model

We will use for simulations the KKW-1 CA-model with a linear dependence of the synchronization distance on the vehicle speed [86]. This KKW CA-model consists of a dynamic part

$$\tilde{v}_{n+1} = \max(0, \min(v_{\text{free}}, v_{\text{s},n}, v_{\text{c},n})) , \tag{33}$$

$$v_{\text{c},n} = \begin{cases} v_n + a\tau & \text{for } g_n > D_n - d, \\ v_n + a\tau \operatorname{sign}(v_{\ell,n} - v_n) & \text{for } g_n \leqslant D_n - d, \end{cases} \tag{34}$$

where $\operatorname{sign}(x)$ is 1 for $x > 0$, 0 for $x=0$ and -1 for $x < 0$, and a stochastic (fluctuation) part

$$v_{n+1} = \max(0, \min(\tilde{v}_{n+1} + a\tau\eta_n, v_n + a\tau, v_{\text{free}}, v_{\text{s},n})) , \tag{35}$$

$$x_{n+1} = x_n + v_{n+1}\tau , \tag{36}$$

$$\eta_n = \begin{cases} -1 & \text{if } r < p_{\text{b}}, \\ 1 & \text{if } p_{\text{b}} \leqslant r < p_{\text{b}} + p_{\text{a}}, \\ 0 & \text{otherwise}, \end{cases} \tag{37}$$

$$p_b(v_n) = \begin{cases} p_0 & \text{if } v_n = 0, \\ p & \text{if } v_n > 0, \end{cases} \tag{38}$$

$$p_a(v_n) = \begin{cases} p_{a1} & \text{if } v_n < v_p, \\ p_{a2} & \text{if } v_n \geqslant v_p. \end{cases} \tag{39}$$

In this KKW CA-model, $n = 0, 1, 2, \ldots$ is number of time steps, τ is the time discretization interval, $\tilde{v}_n$ is the vehicle speed at time step n without fluctuating part, v_n is the vehicle speed at time step n, $v_{\ell,n}$ is the speed of the leading vehicle at time step n, $v_{s,n} = g_n/\tau$ is the safe speed at time step n, v_{free} is the maximal speed (free flow), x_n is the vehicle position at the time step n, $x_{\ell,n}$ is the position of the leading vehicle at time step n, d is the vehicle length which is the same for all vehicles, $g_n = x_{\ell,n} - x_n - d$ is the gap (front to end distance) at time step n, $D_n = d + kv_n\tau$ is the synchronization distance at time step n, η_n is the speed fluctuation at time step n, a is the vehicle acceleration, p_a is the probability of vehicle acceleration, p_b is the probability of vehicle deceleration, r is a random number uniformly distributed between 0 and 1, k, p_0, p, p_{a1}, p_{a2}, v_p are constant parameters.

The steady states for the model are related to a 2D region in the flow–density plane between the line F for free flow ($v = v_{free}$), the line U determined by the safe speed $v_{s,n}$ and the line L determined by the synchronization distance D (Fig. 17(a)). The physics of the KKW CA-model has been considered in Ref. [86]. All numerical simulations below have been performed for a one-lane road with an on-ramp. Models of the road and of the on-ramp are the same and they have the same model parameters (in particular, the road length, the conditions for vehicle squeezing from the on-ramp to the main road, the time and the space discretization units, etc.) as it has been chosen in Ref. [86] for the KKW-1 CA-model, parameter-set I (see Table III in Ref. [86], the KKW-1 CA-model, parameter-set I).

5.4.2. *Transformations of congested patterns at on-ramps*

A diagram of congested pattern at the on-ramp for the KKW CA-model [86] is qualitatively related to the diagram in Fig. 16(a) first predicted within the three-phase traffic theory [76] and then found in the continuum model by Kerner and Klenov [85] (Fig. 17(b)). At the boundary $F_S^{(B)}$ with the probability $P_{FS}^{(B)} = 1$ the F $\rightarrow$ S transition occurs during the time T_{ob}=30 min [86]. Between the boundaries $F_S^{(B)}$ and $S_J^{(B)}$ different SP occurs. Right of the boundary $S_J^{(B)}$ wide moving jams emerge in synchronized flow upstream of the on-ramp, i.e., different GP occur.

To study the congested pattern capacity, in the diagram three lines (Lines 1, 2 and 3) are shown, in addition (Fig. 17(b)). Different congested patterns have been studied when the flow rates q_{on} (Lines 1 and 2) or the flow rate q_{in} (Line 3) are increasing along the related lines.

The transformation of congested patterns along Line 1 is related to a given *high* flow rate in free flow upstream of the on-ramp q_{in} when the flow rate to the on-ramp q_{on} is increased from lower to higher values (Fig. 18). Right of the boundary $F_S^{(B)}$ a WSP occurs (Fig. 18(a)). If the flow rate to the on-ramp is only slightly increased

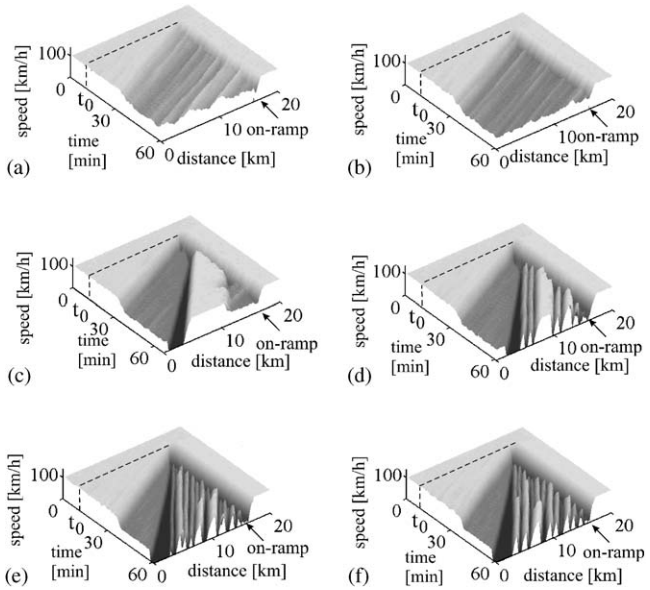

Fig. 18. Evolution of the vehicle speed in space and in time at the given $q_{in} = 2400$ vehicles/h for different q_{on} related to Line 1 in Fig. 17(b): (a,b) widening synchronized flow patterns WSP; (c) the dissolving general pattern (DGP); (d,e,f) general patterns (GP). The flow rate q_{on} is: (a) (40), (b) (60), (c) (105), (d) (150), (e) (200), (f) (500) vehicles/h. The on-ramp is at the location $x = 16$ km (see for more detail Ref. [86]).

then a WSP remains (Fig. 18(b)); however, the vehicle speed in this WSP on average decreases in comparison with the speed in the WSP in Fig. 18(a).

Right of the boundary $S_J^{(B)}$ at the high flow rate q_{in}, which is related to Line 1, a DGP occurs (Fig. 18(c)): After the first wide moving jam has emerged in synchronized flow, the flow rate upstream of the on-ramp decreases because it is now determined by the jam outflow. The maximum flow rate in the wide moving jam outflow is reached when free flow is formed downstream of the jam, $q_{out} = 1810$ vehicles/h. When due to the jam upstream propagation the wide moving jam is far away from the on-ramp an effective flow rate upstream of the on-ramp is determined by q_{out}, i.e., the effective flow rate $q_{in}^{(eff)} = q_{out} = 1810$ vehicles/h. This flow rate is however considerably lower than the initial flow rate $q_{in} = 2400$ vehicles/h. As a result, no wide moving jams can emerge upstream of the on-ramp any more: the DGP occurs which consists of the only one wide moving jam propagating upstream and a SP at the on-ramp. At the mentioned effective flow rate $q_{in}^{(eff)} = 1810$ vehicles/h and the flow rate to the on-ramp $q_{on} = 105$ vehicles/h this SP at the on-ramp is a WSP (Figs. 17(b) and 18(c)). If the flow rate q_{on} is further increased then the pinch region in synchronized flow upstream of the on-ramp appears where narrow moving jams continuously emerge (Fig. 18(d)). Some of these jams transform into wide moving jams leading to the GP formation. This occurs right of the boundary G where GPs are realized. When the flow rate q_{on}

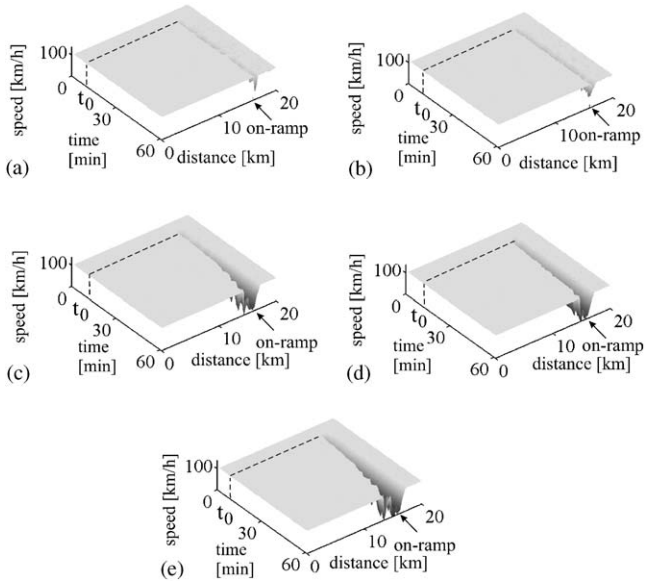

Fig. 19. Evolution of the vehicle speed in space and in time at the given $q_{in} = 1255$ vehicles/h for different q_{on} related to Line 1 in Fig. 17(b): (a,b) localized synchronized flow patterns LSP; (c–e) general patterns (GP). The flow rate q_{on} is: (a) (550), (b) (630), (c) (700), (d) (850), (e) (1000) vehicles/h.

is further increased, the GP does not transform into another pattern: It remains to be a GP at any possible flow rate q_{on} (Fig. 18(e,f)).

The transformation of congested patterns along Line 2 in Fig. 17(b) is related to a given *low* flow rate in free flow upstream of the on-ramp q_{in} when the flow rate to the on-ramp q_{on} is increased beginning from a relative high initial value (Fig. 19). In this case, first a LSP occurs (Fig. 19(a,b)). If the flow rate q_{on} increases then the LSP transforms into a GP (Fig. 19(c–e)). However, whereas for the GP related to Line 1 (Fig. 18(d–f)) the condition $q_{in} > q_{out}$ is fulfilled, in the case of Line 2 we have $q_{in} < q_{out}$ (Fig. 17(b)). This case has already been considered in Ref. [86] where we could see that the most upstream wide moving jam in the GP dissolves over time. This indeed occurs for the GP related to Line 2. For this reason, the width of the GP in Fig. 19(c–e)) increases much slower over time in comparison with the width of the GP in Fig. 18(d–f).

The transformation of congested patterns along Line 3 in Fig. 17(b) is related to a given flow rate to the on-ramp q_{on} when the flow rate in free flow upstream of the on-ramp q_{in} is increased beginning from a relatively low initial value (Fig. 20). Because at this flow rate the boundary $F_S^{(B)}$ is intersected above the boundary W (which separates the WSP and the LSP in the diagram of congested patterns in Fig. 17(b)), a WSP occurs (Fig. 20(a)). The WSP remains in a wide range of the flow rate q_{in} when this flow rate increases (Fig. 20(b,c)). However, the average speed in the WSP decreases whereas the flow rate q_{in} increases. Finally, if the flow rate q_{in} increases and

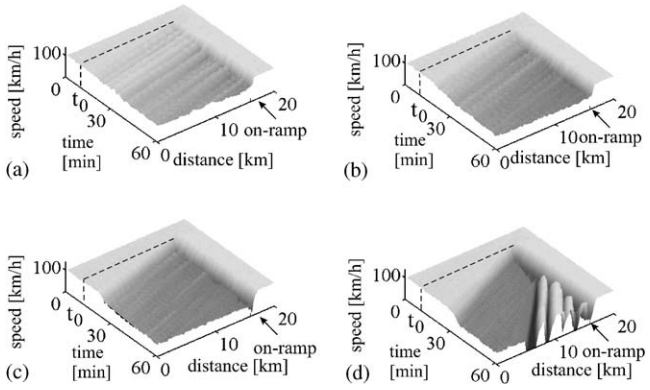

Fig. 20. Evolution of the vehicle speed in space and in time at the given $q_{on} = 200$ vehicles/h for different q_{in} related to Line 3 in Fig. 17(b): (a–c) WSP; (d) GP. The flow rate q_{in} is: (a) (1660), (b) (1800), (c) (1960), (d) (2200) vehicles/h.

the boundary $S_J^{(B)}$ is intersected, a GP occurs (Fig. 20(d)). This GP transforms in a GP shown in Fig. 18(e) when the flow rate q_{in} is further increased.

5.4.3. Time-evolution of discharge flow rate

By the simulation of congested patterns considered above the flow rate to the on-ramp q_{on} has been switched on only after the time $t = t_0$. Let us designate the flow rate far enough downstream of the on-ramp where free flow occurs as q_{down}. During the time $0 \leqslant t < t_0$ free flow occurs at the on-ramp and $q_{down} = q_{sum} = q_{in}$. At $t > t_0$, i.e., after the on-ramp has been switched on, the flow rate q_{down} should increase because $q_{sum} = q_{in} + q_{on}$ at $t \geqslant t_0$. However, at $t \geqslant t_0$ a congested pattern has begun to form upstream of the on-ramp. Thus, at $t > t_0$ the flow rate q_{down} is determined by the discharge flow rate $q_{out}^{(bottle)}$: $q_{down} = q_{out}^{(bottle)}$. The numerical simulation allows us to study the time-evolution of the flow rate q_{down} by the formation of each of the congested patterns upstream of the on-ramp (Figs. 21–23).

This evolution for congested patterns which appear along Line 1 is shown in Fig. 21. It can be seen that although at $t = t_0$ the inflow from the on-ramp is switched on, i.e., as additional vehicles enter the main road from the on-ramp, the discharge flow rate is lower than the flow rate $q_{sum} = q_{in}$ at $0 \leqslant t < t_0$. This is the result of the congested pattern formation: Each of the congested patterns along Line 1 leads to a decrease in the flow rate on the main road just upstream of the on-ramp. This decrease is higher than the increase of the flow rate due to the vehicles squeezing to the main road from the on-ramp.

However, the flow rate q_{down} is considerably lower than q_{sum} when a GP occurs at the bottleneck. When the WSP shown in Fig. 18(a) occurs, then the flow rate q_{down} is only slightly lower than the flow rate q_{sum} (Fig. 21(a)). This means that SPs can be more favorable than GPs in terms of the discharge volume from the congested

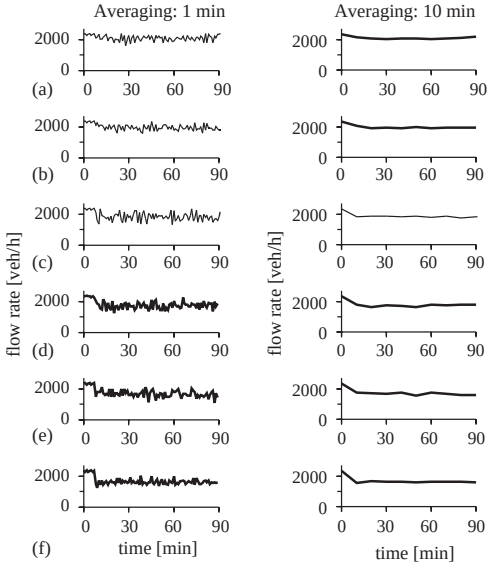

Fig. 21. Time-evolution of the flow rate downstream of the on-ramp q_{down} during the pattern formation at the on-ramp for different congested patterns related to Line 1 in Figs. 17(b) and 18: The flow rate q_{down} averaged during 1 min (left) and the flow rate q_{down} averaged 10 min (right). (a–f) are related to the congested patterns with the same letters (a–f) in Fig. 18. The data from a virtual detector located at $x = 17$ km in free flow downstream of the on-ramp (16 km).

pattern, i.e., in terms of the discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ and of the vehicle delay time due to congestion. The latter is because the average vehicle speed inside the WSP is considerably higher than the average speed in the pinch region of the GP (Fig. 18(d–f)). These conclusions of the theory of congested patterns at bottlenecks are used in methods of congested pattern control at bottlenecks [123].

A different situation is realized for the congested patterns which appear along Line 2 (Fig. 22). In this case, the discharge flow rate is higher than the flow rate $q_{\text{sum}} = q_{\text{in}}$ at $0 \leqslant t < t_0$: Each of the congested patterns along Line 2 leads also to a decrease in the flow rate on the main road just upstream of the on-ramp. However, this decrease is lower than the increase of the flow rate due to the vehicles squeezing onto the main road from the on-ramp.

An intermediate case has been found for the congested patterns which appear along Line 3 (Fig. 23), i.e., when the initial flow rate upstream of the on-ramp q_{in} is changing. First, there is a slight increase in the flow rate $q_{\text{down}} = q_{\text{out}}^{(\text{bottle})}$ due to the congested pattern formation in comparison with the initial flow rate $q_{\text{down}} = q_{\text{in}}$ (Fig. 23(a)). For a higher flow rate q_{in} there is almost no change in the flow rate q_{down} after congested patterns have been formed (Fig. 23(b,c)). When the flow rate q_{in} is further increased, the discharge flow rate is lower than the initial flow rate $q_{\text{down}} = q_{\text{in}}$ at $0 \leqslant t < t_0$ as it was for the congested patterns along Line 1 (Fig. 23(d)), i.e., the flow rate q_{down} decreases during the congested pattern formation.

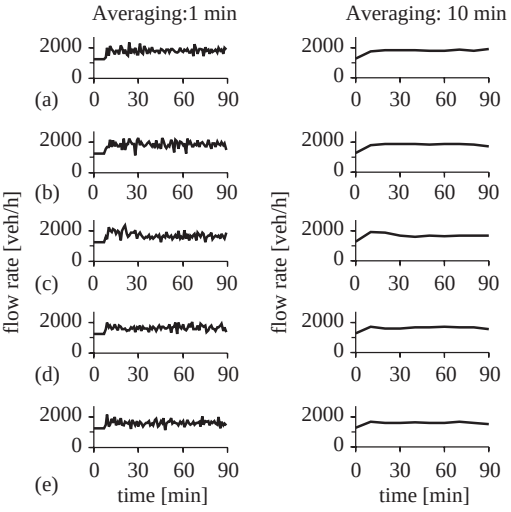

Fig. 22. Time-evolution of the flow rate downstream of the on-ramp q_{down} during the pattern formation at the on-ramp for different congested patterns related to Line 2 in Figs. 17(b) and 19: The flow rate q_{down} averaged during 1 min (left) and the flow rate q_{down} averaged 10 min (right). (a–e) are related to the congested patterns with the same letters (a–e) in Fig. 19. The data from a virtual detector located at $x = 17$ km in free flow downstream of the on-ramp (16 km).

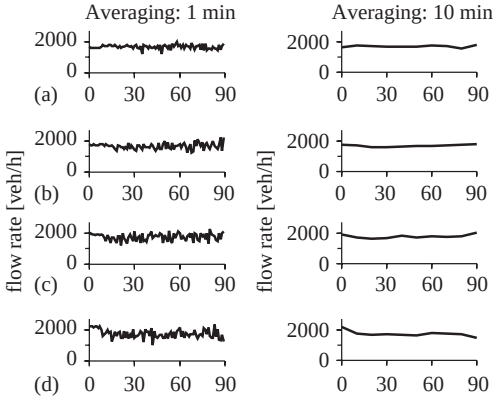

Fig. 23. Time-evolution of the flow rate downstream of the on-ramp q_{down} during the pattern formation at the on-ramp for different congested patterns related to the Line 3 in Figs. 17(b) and 20: The flow rate q_{down} averaged during 1 min (left) and the flow rate q_{down} averaged 10 min (right). (a–d) are related to the congested patterns with the same letters (a–d) in Fig. 20. The data from a virtual detector located at $x = 17$ km in free flow downstream of the on-ramp (16 km).

5.4.4. Congested pattern capacity

The congested pattern capacity $q_{\text{cong}}^{(\text{B})}$ (28) has been calculated through the 60 min averaging of the discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ after the related congested pattern has been

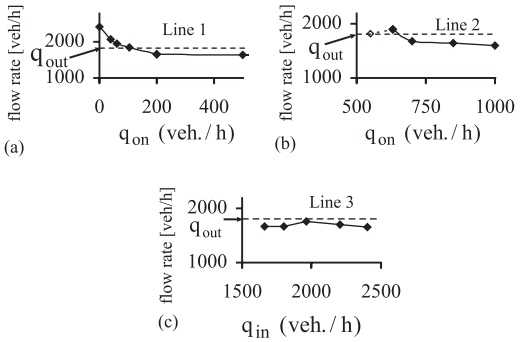

Fig. 24. Dependence of the average congested pattern capacity on the pattern type and pattern parameters: (a,b). The congested pattern capacity as function on the flow rate q_{on} related to Line 1 (a) and Line 2 (b) in Fig. 17(b). (c) The congested pattern capacity as function on the flow rate q_{in} related to Line 3 in Fig. 17(b). The black points correspond to the congested patterns in Figs. 18–20 for each of Lines 1–3, respectively. Dashed lines show the flow rate in the wide moving jam outflow, q_{out}. The averaging of the discharge follow rate during 60 min beginning from the time moment $t = 20$ min (after congested patterns have been formed) is performed. The data from a virtual detector located at $x = 17$ km in free flow downstream of the on-ramp (16 km).

formed. It has been found that the congested pattern capacity $q_{cong}^{(B)}$ depends on the type of congested patterns and on the pattern parameters noticeably.

Indeed, along Line 1 (Fig. 24(a)) the congested pattern capacity $q_{cong}^{(B)}$ is a decreasing function of the flow rate to the on-ramp q_{on}. This is linked to the pinch effect in synchronized flow which occurs when q_{on} increases and a GP is formed. This leads to a decrease in the flow rate in synchronized flow upstream of the on-ramp. The decrease in the latter flow rate is noticeably higher than an increase in the flow rate to the on-ramp. For this reason, the congested pattern capacity $q_{cong}^{(B)}$ decreases although q_{on} increases. The more the decrease in the speed and in the flow rate in synchronized flow upstream of the on-ramp is the higher the flow rate q_{on} is. However, there is a saturation of the decrease in the congested pattern capacity $q_{cong}^{(B)}$ when q_{on} is further increased. The congested pattern capacity at lower q_{on} is higher than the flow rate in the wide moving jam outflow, q_{out} (dashed line in Fig. 24(a)). At higher q_{on} the congested pattern capacity becomes lower than q_{out}.

By calculation of the congested pattern capacity condition (27) has been fulfilled. The only exception is the calculation of the capacity along Line 2: In this case, first a LSP appears (Fig. 19(a,b)) and therefore the condition (32) has been used. The congested pattern capacity which is related to this LSP has been determined by the maximum discharge flow rate at which the LSP still exists upstream of the on-ramp. It has been found that the latter condition is only fulfilled for the LSP at $q_{on} \approx 650$ vehicles/h. This has been taken into account in Fig. 24(b): The dashed curve is related to the discharge flow rate from the LSP when the capacity is not reached and the solid curve corresponds to the congested pattern capacity. It can been seen that the congested pattern capacity decreases when the flow rate q_{on} increases. However, in comparison with Line 1 the decrease in the capacity along Line 2 is considerably lower at the same increase in the flow rate q_{on}.

Along Line 3 (Fig. 24(c)) the congested pattern capacity $q_{\rm cong}^{\rm (B)}$ is a non-monotonous function of the flow rate to the on-ramp $q_{\rm on}$: There is a maximum point on this dependence. It occurs that at the same flow rate to the on-ramp $q_{\rm on}$ the flow rate in synchronized flow in the WSP upstream of the on-ramp increases during the evolution of the WSP (from the WSP in Fig. 20(a) to the WSP in Fig. 20(c)) with the increase in the flow rate $q_{\rm in}$. The flow rate in synchronized flow of the WSP in the maximum point in Fig. 24(c) (the WSP in Fig. 20(c)) is higher than the flow rate in the wide moving jam outflow $q_{\rm out}$ (dashed line in Fig. 24(c)). When the flow rate $q_{\rm in}$ is further increased and therefore the boundary $S_{\rm J}^{\rm (B)}$ is intersected, then wide moving jams begin to form in synchronized flow of the initial WSP: a GP is forming. The flow rate in the wide moving jam outflow cannot exceed $q_{\rm out}$. As a result, the flow rate through synchronized flow of the GP occurs to be noticeably lower than in the WSP in Fig. 20(c) which is related to the maximum point of the congested pattern capacity in Fig. 24(c). Thus, the congested pattern capacity decreases when the WSP transforms into the GP. This explains the maximum point in the congested pattern capacity as function of the flow rate $q_{\rm in}$.

6. Conclusions

The three-phase traffic theory by the author describes phase transitions and a diverse variety of spatial–temporal congested patterns both on homogeneous roads and at highway bottlenecks which are related to results of empirical observations [74,76]. The features of these phase transitions and of the spatial–temporal congested patterns at bottlenecks in the three-phase traffic theory [74,76,12,80,84,85,101,103] are qualitatively different in comparison with the related results which have been derived in the fundamental diagram approach [31–34]. An exception is only the propagation of wide moving jams whose characteristic parameters and features appear to play an important role (in particular, the flow rate in the outflow from the jam, $q_{\rm out}$) in both the three-phase traffic theory and traffic flow theories in the fundamental diagram approach.

Recent empirical results of a study of the congested patterns at on- and off-ramps and of their evolution when the bottleneck strength is gradually changing [76] confirm the discussed results and conclusions of the three-phase-traffic-theory [74,76,12,80,84,85,103] rather than the related results and conclusions of traffic theories in the fundamental diagram approach [31–34,62–66].

The three-phase traffic theory [78,74,80,84,85,103] has also been confirmed by the on-line application in the traffic center of the State Hessen of some recent models “ASDA” (Automatische Staudynamikanalyse: Automatic Tracing of Moving Traffic Jams) and “FOTO” (Forecasting of Traffic Objects) [115–122] which are based on this traffic flow theory. These models allow reconstruction, tracing and prediction of spatial–temporal traffic dynamics based on local measurements of traffic. The models ASDA and FOTO perform without validation of model parameters at different traffic conditions (see a recent review about the models ASDA and FOTO in Ref. [105]).

Mathematical microscopic traffic flow models in the frame of the three-phase traffic theory which have recently been proposed by Kerner and Klenov [85,87] shows a considerable potential both for the development of qualitatively new mathematical traffic models and for the traffic flow theory development on the basis of the discussed hypotheses of the three-phase traffic theory. This is also confirmed by results of a numerical study of several new cellular automata traffic flow models which have recently been developed by Kerner et al. [86].

Based on the three-phase traffic theory, a general probabilistic theory of highway capacity has been developed and presented in this paper. It is shown that already in free traffic at the on-ramp there may be an infinite multitude of highway capacities which depend on the flow rate to the on-ramp. When a bottleneck is congested then a much more complicated picture of congested pattern highway capacity can be realized: Congested pattern highway capacity strongly depends on the type of the congested pattern and the pattern parameters. The theory of highway capacity and of the capacity drop presented in the paper is confirmed by both results of empirical observations [10,11,79,76] and by the presented numerical results of the simulations of the KKW CA-model presented in the article and in Ref. [86].

As it has recently been shown in Ref. [123], the microscopic three-phase traffic flow theory [85–87] is also the basis for a study of different traffic control strategies and for an analysis of driver assistance systems. A detailed consideration of the physics of traffic (features of empirical spatial–temporal traffic patterns, traffic flow theory and engineering applications) has been made in the book [124].

Acknowledgements

I would like to thank Hubert Rehborn and Sergey Klenov for their help and acknowledges funding by BMBF within project DAISY.

References

[1] M. Koshi, M. Iwasaki, I. Ohkura, in: V.F. Hurdle, et al., (Eds.), Proceedings of 8th International Symposium on Transportation and Traffic Theory, University of Toronto Press, Toronto, Ontario, 1983, p. 403.
[2] I. Prigogine, R. Herman, Kinetic Theory of Vehicular Traffic, American Elsevier, New York, 1971.
[3] J. Treiterer, Investigation of traffic dynamics by aerial photogrammetry techniques, Ohio State University Technical Report PB 246 094, Columbus, OH, 1975.
[4] P. Athol, A. Bullen, Highway Research Record, Vol. 456, HRB, National Research Council, Washington, DC, 1973, pp. 50–54.
[5] J.H. Banks, Transp. Res. Rec. 1287 (1990) 20–28.
[6] J.H. Banks, Transp. Res. Rec. 1320 (1991) 83–90.
[7] F.L. Hall, K. Agyemang-Duah, Transp. Res. Rec. 1320 (1991) 91–98.
[8] F.L. Hall, V.F. Hurdle, J.H. Banks, Transp. Res. Rec. 1365 (1992) 12–18.
[9] L. Elefteriadou, R. Roess, W. McShane, Transp. Res. Rec. 1484 (1995) 80–89.
[10] B.S. Kerner, H. Rehborn, Phys. Rev. Lett. 79 (1997) 4030.
[11] B. Persaud, S. Yagar, R. Brownlee, Transp. Res. Rec. 1634 (1998) 64.
[12] B.S. Kerner, in: A. Ceder (Ed.), Transportation and Traffic Theory, Elsevier, Amsterdam, 1999, pp. 147–171.

[13] C.F. Daganzo, Fundamentals of Transportation and Traffic Operations, Elsevier, New York, 1997.
[14] A.D. May, Traffic Flow Fundamental, Prentice-Hall, Englewood Cliffs, NJ, 1990.
[15] Highway Capacity Manual, Transportation Research Boad, Washington, DC, 2000.
[16] B.S. Kerner, Dependence of empirical fundamental diagram on spatial-temporal traffic patterns features, cond-mat/0309018, e-print in http://arxiv.org/abs/cond-mat/0309018.
[17] M.J. Lighthill, G.B. Whitham, Proc. R. Soc. A 229 (1955) 317.
[18] P.I. Richards, Oper. Res. 4 (1956) 42.
[19] R. Herman, E.W. Montroll, R.B. Potts, R. W. Rothery, Oper. Res. 7 (1959) 86–106.
[20] D.C. Gazis, R. Herman, R.W. Rothery, Oper. Res. 9 (1961) 545–567.
[21] E. Kometani T. Sasaki, J. Oper. Res. Soc. Jpn. 2 (1958) 11;
E. Kometani T. Sasaki, Oper. Res. 7 (1959) 704;
E. Kometani T. Sasaki, in: R. Herman (Ed.), Theory of Traffic Flow, Elsevier, Amsterdam, 1961, p. 105.
[22] G.F. Newell, Oper. Res. 9 (1961) 209.
[23] I. Prigogine, in: R. Herman (Ed.), Theory of Traffic Flow, Elsevier, Amsterdam, 1961, p. 158.
[24] H.J. Payne, in: G.A. Bekey (Ed.), Mathematical Models of Public Systems, Vol. 1, Simulation Council, La Jolla, 1971.
[25] P.G. Gipps, Trans. Res. B. 15 (1981) 105–111.
[26] W. Leutzbach, Introduction to the Theory of Traffic Flow, Springer, Berlin, 1988.
[27] R. Wiedemann, Simulation des Verkehrsflusses, University of Karlsruhe, Karlsruhe, 1974.
[28] G.B. Whitham, Linear and Nonlinear Waves, Wiley, New York, 1974.
[29] M. Cremer, Der Verkehrsfluss auf Schnellstrassen, Springer, Berlin, 1979.
[30] G.F. Newell, Applications of Queuing Theory, Chapman & Hall, London, 1982.
[31] D. Chowdhury, L. Santen, A. Schadschneider, Phys. Rep. 329 (2000) 199.
[32] D. Helbing, Rev. Mod. Phys. 73 (2001) 1067–1141.
[33] T. Nagatani, Rep. Prog. Phys. 65 (2002) 1331–1386.
[34] K. Nagel, P. Wagner, R. Woesler, Oper. Res. (2003)
[35] J.-B. Lesort (Ed.), Transportation and traffic theory, Proceedings of the 13th International Symposium on Transportation and Traffic Theory, Elsevier, Oxford, 1996.
[36] A. Ceder (Ed.), Transportation and traffic theory, Proceedings of the 14th International Symposium on Transportation and Traffic Theory, Elsevier, Oxford, 1999.
[37] M.A.P. Taylor (Ed.), Transportation and traffic theory in the 21st Century, Proceedings of the 15th International Symposium on Transportation and Traffic Theory, Elsevier, Oxford, 2002.
[38] D.E. Wolf, M. Schreckenberg, A. Bachem (Eds.), Traffic and granular flow, Proceedings of the International Workshop on Traffic and Granular Flow, October 1995, World Scientific, Singapore, 1995.
[39] M. Schreckenberg, D.E. Wolf (Eds.), Traffic and granular flow' 97, Proceedings of the International Workshop on Traffic and Granular Flow, October 1997, Springer, Singapore, 1998.
[40] D. Helbing, H.J. Herrmann, M. Schreckenberg, D.E. Wolf (Eds.), Traffic and granular flow' 99, Proceedings of the International Workshop on Traffic and Granular Flow, October 1999, Springer, Heidelberg, 2000.
[41] M. Fukui M, Y. Sugiyama, M. Schreckenberg, D.E. Wolf (Eds.), Traffic and granular flow' 01, Proceedings of the International Workshop on Traffic and Granular Flow, October 2001, Springer, Heidelberg, 2003.
[42] K. Nagel, M. Schreckenberg, J Phys. (France) I 2 (1992) 2221.
[43] K. Nagel, M. Paczuski, Phys. Rev. E 51 (1995) 2909.
[44] M. Schreckenberg, A. Schadschneider, K. Nagel, N. Ito, Phys. Rev. E 51 (1995) 2939.
[45] G.B. Whitham, Proc. R. Soc. London A 428 (1990) 49.
[46] B.S. Kerner, P. Konhäuser, Phys. Rev. E 50 (1994) 54–83.
[47] M. Bando, K. Hasebe, A. Nakayama, A. Shibata, Y. Sugiyama, Phys. Rev. E 51 (1995) 1035–1042.
[48] M. Bando, K. Hasebe, A. Nakayama, A. Shibata, Y. Sugiyama, J. Phys. I France 5 (1995) 1389.
[49] Y. Sugiyama, H. Yamada, Phys. Rev. E 55 (1997) 7749.
[50] M. Herrmann, B.S. Kerner, Physica A 255 (1998) 163–188.
[51] S. Krauß, P. Wagner, C. Gawron, Phys. Rev. E 53 (1997) 5597.

[52] R. Barlovic, L. Santen, A. Schadschneider, M. Schreckenberg, Eur. Phys. J. B. 5 (1998) 793.
[53] T. Nagatani, Phys. Rev. E 58 (1998) 4271–4279;
T. Nagatani, Phys. Rev. E 59 (1999) 4857–4864;
T. Nagatani, Phys. Rev. E 60 (1999) 180–187;
T. Nagatani, Phys. Rev. E 60 (1999) 6395–6401;
T. Nagatani, Phys. Rev. E 61 (2000) 3564–3570;
T. Nagatani, Phys. Rev. E 61 (2000) 3534–3540;
T. Nagatani, Phys. Rev. E 64 (2001) 036 115;
T. Nagatani, Phys. Rev. E 64 (2001) 016 106;
T. Nagatani, Physica A 290 (2001) 501–511;
T. Nagatani, Physica A 280 (2000) 602–613.
[54] T. Nagatani, Physica A 258 (1998) 237.
[55] D. Helbing, M. Schreckenberg, Phys. Rev. E 59 (1999) R2505.
[56] M. Treiber, A. Hennecke, D. Helbing, Phys. Rev. E 59 (1999) 239.
[57] R. Mahnke, J. Kaupužs, Phys. Rev. E 59 (1999) 117.
[58] B.S. Kerner, S.L. Klenov, P. Konhäuser, Phys. Rev. E 56 (1997) 4200–4216.
[59] B.S. Kerner, P. Konhäuser, M. Schilke, Phys. Rev. E 51 (1995) 6243–6246.
[60] H.Y. Lee, H.-W. Lee, D. Kim, Phys. Rev. Lett. 81 (1998) 1130.
[61] D. Helbing, M. Treiber, Phys. Rev. Lett. 81 (1998) 3042–3045.
[62] D. Helbing, A. Hennecke, M. Treiber, Phys. Rev. Lett. 82 (1999) 4360.
[63] H.Y. Lee, H.-W. Lee, D. Kim, Phys. Rev. E 59 (1999) 5101.
[64] H.Y. Lee, H.-W. Lee, D. Kim, Physica A 281 (2000) 78.
[65] H.Y. Lee, H.-W. Lee, D. Kim, Phys. Rev. E 62 (2000) 4737.
[66] M. Treiber, A. Hennecke, D. Helbing, Phys. Rev. E 62 (2000) 1805.
[67] M. Treiber, D. Helbing, J. Phys. A: Math. Gen. 32 (1999) L17.
[68] E. Tomer, L. Safonov, S. Havlin, Phys. Rev. Lett. 84 (2000) 382.
[69] W. Knospe, L. Santen, A. Schadschneider, M. Schreckenberg, J. Phys. A 33 (2000) L477.
[70] W. Knospe, L. Santen, A. Schadschneider, M. Schreckenberg, Phys. Rev. E 65 (2002) 015101(R).
[71] D. Helbing, D. Batic, M. Schönhof, M. Treiber, cond-mat/0108548 (2001);
D. Helbing, D. Batic, M. Schönhof, M. Treiber, Physica A 303 (2002) 251–260.
[72] E. Tomer, L. Safonov, N. Madar, S. Havlin, cond-mat/0105493 (2001);
E. Tomer, L. Safonov, N. Madar, S. Havlin, Phys. Rev. E 65 (2002) 065101(R).
[73] M. Treiber, D. Helbing, cond-mat/0304337 (2003);
M. Treiber, D. Helbing, Phys. Rev. E 68 (2003).
[74] B.S. Kerner, Phys. Rev. Lett. 81 (1998) 3797.
[75] B.S. Kerner, J. Phys. A: Math. Gen. 33 (2000) L221.
[76] B.S. Kerner, Phys. Rev. E 65 (2002) 046 138.
[77] B.S. Kerner, H. Rehborn, Phys. Rev. E 53 (1996) R4275–R4278.
[78] B.S. Kerner, in: R. Rysgaard (Ed.), Proceedings of the 3rd Symposium on Highway Capacity and Level of Service, Vol. 2, Road Directorate, Ministry of Transport, Denmark, 1998, pp. 621–642.
[79] B.S. Kerner, Transp. Res. Rec. 1710 (2000) 136–144.
[80] B.S. Kerner, Transp. Res. Rec. 1678 (1999) 160–167.
[81] B.S. Kerner, in: D. Helbing, H.J. Herrmann, M. Schreckenberg, D.E. Wolf (Eds.), Traffic and granular flow' 99, Proceedings of the International Workshop on Traffic and Granular Flow, October 1999, Springer, Heidelberg, 2000, pp. 253–284.
[82] B.S. Kerner, Networks and Spatial Econ. 1 (2001) 35.
[83] B.S. Kerner, in: M. Schreckenberg, D.E. Wolf (Eds.), Traffic and granular flow' 97, Proceedings of the International Workshop on Traffic and Granular Flow, October 1997, Springer, Singapore, 1998, pp. 239–267.
[84] B.S. Kerner, Phys. World 12 (8) (1999) 25–30.
[85] B.S. Kerner, S.L. Klenov, J. Phys. A: Math. Gen. 35 (2002) L31.
[86] B.S. Kerner, S.L. Klenov, D.E. Wolf, cond-mat/0206370 (2002);
B.S. Kerner, S.L. Klenov, D.E. Wolf, J. Phys. A: Math. Gen. 35 (2002) 9971–10013.
[87] B.S. Kerner, S.L. Klenov, Phys. Rev. E 68 (2003) 036 130.

[88] E. Schöll, Nonequilibrium Phase Transitions in Semiconductors, Reidel, Dordrecht, 1987.
[89] V.A. Vasil'ev, Yu.M. Romanovskii, D.S. Chernavskii, V.G. Yakhno, Autowave Processes in Kinetic Systems, Springer, Berlin, 1990.
[90] A.S. Mikhailov, Foundations of Synergetics, 2nd Edition, Vol. I, Springer, Berlin, 1994.
[91] B.S. Kerner, V.V. Osipov, Autosolitons: A New Approach to Problems of Self-organization and Turbulence, Kluwer, Dordrecht, Boston, London, 1994.
[92] K. Nishinari, D. Takahashi, J. Phys. A 31 (1998) 5439;
K. Nishinari, D. Takahashi, J. Phys. A 32 (1999) 93;
K. Nishinari, J. Phys. A 34 (2001) 10 723.
[93] J. Matsukidaira, K. Nishinari, Phys. Rev. Lett. 980 (2003) 088 701.
[94] M. Fukui, K. Nishinari, D. Takahashi, Y. Ishibashi, Physica A 303 (2002) 226–238.
[95] R. Kühne, R. Mahnke, I. Lubashevsky, J. Kaupužs, Phys. Rev. E 65 (2002) 066 125.
[96] D. Jost, K. Nagel, 2002; cond-mat/0208082.
[97] I. Lubashevsky, R. Mahnke, Phys. Rev. E 62 (2000) 6082.
[98] I. Lubashevsky, R. Mahnke, P. Wagner, S. Kalenkov, Phys. Rev. E 66 (2002) 016 117.
[99] P. Nelson, Phys. Rev. E 61 (2000) R6052.
[100] S. Kriso, R. Friedrich, J. Peinke, P. Wagner, cond-mat/0110084 (2001).
[101] B.S. Kerner, in: M. Fukui M, Y. Sugiyama, M. Schreckenberg, D.E. Wolf (Eds.), Traffic and granular flow' 01, Proceedings of the International Workshop on Traffic and Granular Flow, October 2001, Springer, Heidelberg, 2003.
[102] D. Helbing, M. Treiber, trafficforum/02031301 (2002)
D. Helbing, M. Treiber, Cooperative Transportation Dynamics, Vol. 1, 2002, 2.1.–2.24.
[103] B.S. Kerner, Transp. Res. Rec. 1802 (2002) 145–154.
[104] B.S. Kerner, Math. Comput. Modelling 35 (2002) 481–508.
[105] B.S. Kerner, H. Rehborn, M. Aleksic, A. Haug, in: M. Schreckenberg, R. Selten (Eds.), Human Behavior and Traffic Networks, Springer, Berlin, 2004.
[106] M. Fukui, Y. Ishibashi, J. Phys. Soc. Japan 65 (1996) 1868.
[107] R. Mahnke, N. Pieret, Phys. Rev. E 56 (1997) 2666.
[108] K. Nishinari, D. Takahashi, J. Phys. A 33 (2000) 7709.
[109] M. Takayasu, H. Takayasu, Fractals 1 (1993) 860.
[110] D.E. Wolf, Physica A 263 (1999) 438.
[111] T. Nagatani, K. Nakanishi, Phys. Rev. E 57 (1998) 6415.
[112] S. Rosswog, P. Wagner, Phys. Rev. E 65 (2002) 036 106.
[113] B.S. Kerner, H. Rehborn, Phys. Rev. E 53 (1996) R1297–R1300.
[114] B.S. Kerner, in: M.A.P. Taylor (Ed.), Transportation and Traffic Theory in the 21th Century, Elsevier, Amsterdam, 2002, pp. 417–439.
[115] B.S. Kerner, German Patent DE 199 44 075.
[116] B.S. Kerner, in: 2001 IEEE Intelligent Transportation Systems Proceedings, IEEE, Oakland, USA, 2001, pp. 88–93.
[117] B.S. Kerner, H. Rehborn, H. Kirschfink, German Patent DE 196 47 127;
B.S. Kerner, H. Rehborn, H. Kirschfink, US-Patent US 5861820 (1998).
[118] B.S. Kerner, H. Rehborn, German Patent Publication DE 198 35 979.
[119] B.S. Kerner, M. Aleksic, U. Denneler, German Patent DE 199 44 077.
[120] B.S. Kerner, H. Rehborn, M. Aleksic, A. Haug, Traffic Eng. Control 42 (2001) 282–287.
[121] B.S. Kerner, H. Rehborn, M. Aleksic, A. Haug, R. Lange, Traffic Eng. Control 42 (2001) 345–350.
[122] B.S. Kerner, H. Rehborn, M. Aleksic, A. Haug, Transp. Res. C (2003), in press.
[123] B.S. Kerner, Control of spatial-temporal congested traffic patterns at highway bottlenecks, cond-mat/0309017, e-print in http//axriv.org/abs/cond-mat/0309017;
B.S. Kerner, in: Proceedings of the 10th World Congress on Intelligent Transport Systems, Paper No. 2043 T, Madrid, Spain, 2003;
B.S. Kerner, in: Proceedings of the 83rd Annual Meeting of Transportation Research Board, TRB Paper No. 04-3062, January 11–15, 2004, Washington, D.C., 2004.
[124] B.S. Kerner, The Physics of Traffic, Springer, Berlin, 2004, in press.