Asia-Pacific Journal of Operational Research
Vol. 24, No. 4 (2007) 435–461

MODELING TRAFFIC FLOWS WITH QUEUEING MODELS: A REVIEW

TOM VAN WOENSEL*
Department of Technology Management, Eindhoven University of Technology
Den Dolech 2, Eindhoven, 5600MB, The Netherlands
t.v.woensel@tm.tue.nl

NICO VANDAELE
Department of Applied Economic Sciences, University of Antwerp
Prinsstraat 13, Antwerp, B2000, Belgium
nico.vandaele@ua.ac.be

Received 3 November 2005
Accepted 9 March 2006

In this paper, an overview of different analytic queueing models for traffic on road networks is presented. In the literature, it has been shown that queueing models can be used to adequately model uninterrupted traffic flows. This paper gives a broad review on this literature. Moreover, it is shown that the developed published methodologies (which are mainly single node oriented) can be extended towards queueing networks. First, an extension towards queueing networks with infinite buffer sizes is evaluated. Secondly, the assumption of infinite buffer sizes is dropped leading to queueing networks with finite buffer sizes. The impact of the buffer size when comparing the different queueing network methodologies is studied in detail. The paper ends with an analytical application tool to facilitate the optimal positioning of the counting points on a highway.

Keywords: Traffic flows; queueing networks; finite and infinite buffer sizes.

1. Introduction and Motivation

Congestion is a function of the number of vehicles on the road, showing the need for well-performing traffic models that capture this specific relationship. Traffic flows are usually modelled empirically: speed and flow data are collected for a specific road and econometrically fitted into curves, i.e., the speed-flow-density diagrams (Daganzo, 1997). Alternatively, (mainly supported by the increasing computer performance), one may use simulation to model traffic flows [e.g., leading to the well-known car-following models, see Transportation Research Board (1996); Zhang and Kim (2005)]. These approaches are however limited in terms of predictive power

*Corresponding author.

and sensitivity analysis. Moreover, these techniques are highly data-dependent and (computer) time-dependent and as such, not directly applicable in the decision process of various policy makers (Jain and MacGregor Smith, 1997).

As an alternative to these methodologies, analytical models based on queueing theory could be used to model traffic flows. This review paper intends to give an overview of the different efforts in the relevant traffic flow literature where queueing models are used. The contributions of this paper are twofold:

First, the methodology to model road networks using analytical queueing models is reviewed in detail. The current methodology is however mostly limited to single node analysis, i.e., single stage queueing models. As in practice traffic passes through a multitude of nodes, the extension towards network models is necessary. In this paper, it is proposed that a road network can be represented as a queueing network where vehicles spend time. This time spent is dependent upon the occupation of the road network, i.e., a high occupation or traffic intensity will lead to more time en route. Consequently, the performance indicators of the queueing networks will be used to determine the time on the road. Note that the term congestion will be used here in a strictly queueing theory sense meaning more than one customer in the system leading to traffic intensity strictly larger than zero. When considering getting stuck in traffic (stand still), the term traffic jam will be used.

Secondly, in the late 1990*s*, more and more vehicle detectors have been installed throughout the world to record the passing of vehicles (Newell, 2002; Ehlert *et al.*, 2005). Mostly, the decision concerning the location of the detector is arbitrarily (e.g., near an off-ramp or on-ramp). Based on the insights obtained from the literature on finite versus infinite queueing networks applied to traffic environments, a policy tool is developed to determine the optimal positions of the vehicle detectors on highways. The tool proposed in this paper determines, based on the expected traffic intensity, the optimal number and the best locations for the different detectors to adequately monitor traffic.

This paper is organized as follows. First, in Sec. 2, a broad literature review on queueing models applied to traffic flows is presented. Based on the latter, an extension in the direction of queueing network analysis for traffic networks is presented. It is split up into two major paths depending upon the buffer size: nodes having an infinite buffer size (Sec. 3.1) or nodes having a finite buffer size (Sec. 3.2). The developed models (networks with infinite and finite buffer size) are compared with each other and differences are evaluated (Sec. 4). In Sec. 5, a tool (based on the elaborated queueing analysis) to determine the optimal places of the different counting points on the road is presented. Then, future research opportunities are discussed (Sec. 6). The last section concludes this review.

2. Literature Overview

In this paper, traffic flow models based on queueing theory are considered. The following subsections explain the basic concepts in detail and give relevant references.

The interested reader on the history of traffic flow theory in general, is referred to, e.g., Newell (2002) or Daganzo (1997).

2.1. *Modeling traffic flows*

Traffic flows can be divided into two primary types: uninterrupted versus interrupted traffic flows (Transportation Research Board, 1996):

(1) The first type, *Uninterrupted flows*, is defined as all the flows regulated by vehicle-vehicle interactions and interactions between vehicles and the roadway. For example, vehicles traveling on a highway are participating in uninterrupted flows.
(2) *Interrupted flow*, the second type of traffic flows, is flow regulated by an external means, such as a traffic signal. Under interrupted flow conditions, vehicle-vehicle interactions and vehicle-roadway interactions play a secondary role in defining the traffic flow.

Understanding what type of flow is occurring in a given situation will lead to different methods for analyzing traffic situations. In this paper, only uninterrupted flows are considered, i.e., only traffic on highways are considered and modeled using queueing models.

In the literature, three main types of model representations are considered depending upon the level of detail: microscopic, mesoscopic and macroscopic models. More specifically, microscopic models deal with vehicles, where macroscopic models deal with flows. The mesoscopic representation combines microscopic and macroscopic elements in a unified approach.

Microscopic models describe each vehicle separately and are based on theories of how vehicles maneuver through traffic. Microscopic models are oriented towards:

- Car-following theories: These models describe the behavior of a vehicle following another vehicle (Transportation Research Board, 1996).
- Time-space diagrams: A time-space diagram is commonly used to solve a number of transportation-related problems. Typically, time is drawn on the horizontal axis and distance from a reference point on the vertical axis. The trajectories of individual vehicles in motion are portrayed in this diagram by sloping lines, and stationary vehicles are represented by horizontal lines. The slope of the line represents the speed of the vehicle. Curved portions of the trajectories represent vehicles undergoing speed changes such as deceleration (Daganzo, 1997).
- Microsimulation: These models try to model and visualize transport systems at the basic level of a vehicle. Software tools that can be used here are, e.g., *AIMSUN*, *MICROSIM*, *PARAMICS*, *VISSIM*, etc. See the internet page http://www.its.leeds.ac.uk/projects/smartest/links.html for more information and a comparison of the different microsimulation models.

In *macroscopic* models, all individual vehicles are aggregated and described as flows. In general, macroscopic models can be categorized into one of the following approaches:

- Capacity analysis: These models try to obtain insights into all aspects of capacity and level-of-service analyses for highway (and other) facilities. More information on these models, can be found in Transportation Research Board (1998).
- Speed-flow-density relationships: These models use the speed-flow-density relationships as a basis. The first model that mathematically captures this relationship is the one developed by Greenshields (1935).
- Shock wave analysis: Shock waves in traffic are very similar to the waves produced by dropping stones in water. A shock wave propagates along a line of vehicles in response to changing conditions at the front of the line. Shock waves can be generated by collisions, sudden increases in speed caused by entering free flow conditions, etc. Basically, a shock wave exists whenever the traffic conditions change.

Mesoscopic models mix elements of the microscopic and macroscopic models in an attempt to get the best of both worlds. Mesoscopic models describe the traffic entities at a high level of detail, but their behavior and interactions are described at a lower level of detail. For instance, a lane-change maneuver might be represented for an individual vehicle as an instantaneous event, where the decision to perform a lane-change is based on, e.g., relative lane densities and speeds obtained via the macroscopic models (see, e.g., Ben-Akiva *et al.*, 1998). Mesoscopic models can be categorized as follows:

- Headway distribution models: These models describe the distribution of the headways (i.e., the passage times difference of two successive vehicles) of the individual vehicles, while neither explicitly considering nor tracing each vehicle separately (Branston, 1976).
- Cluster models: A cluster is a group of vehicles that share a specific property. These clusters are then routed through the network and act as one entity. The speeds on each road is derived from a speed-density function defined for that road (Leonard *et al.*, 1989).
- Gas-kinetic continuum models: These models describe the dynamics of velocity distributions (Hongler and Filliger, 2002). Gas-kinetic models implicitly bridge the gap [via, e.g., Boltzmann approaches as in the seminal paper by Prigogine and Andrews (1960)] between microscopic driver behavior and the aggregated macroscopic modeling approach so that more complex and nonlinear dynamics can be reproduced (Nelson and Sopasakis, 1998).
- Mesoscopic simulation (De Palma and Marchal, 2002): The main application area of mesoscopic simulation models is when the detail of a microscopic simulation might be desirable but infeasible due to the large network to be analyzed.

The main disadvantages of microscopic and mesoscopic models are their complexity, their non-analytical character and their large demand on computer time. These models are more simulations tools rather than mathematical models. Due to these drawbacks, the usability of microscopic and mesoscopic models is mostly limited to sections of roads, rather than networks. Therefore, macroscopic models are more suitable for the design of control strategies since they describe the traffic flow process analytically and demand lower computational time (May, 1990).

2.2. *Speed-flow-density diagrams*

It is usually observed that the speed for a certain time period tends to be reproduced whenever the same flow is observed. Based on this observation, it seems reasonable to postulate that, if traffic conditions on a given road are stationary, there should be a relationship between flow, speed, and density. The specific relationship between speed, flow and density results in the concept of speed-flow-density diagrams, which describe the relationships between traffic flow (q), density (k) and speed (v). The seminal work on speed-flow diagrams was the paper by Greenshields (1935). In this paper, he derived a parabolic expression for the speed-flow diagram. Over the years, the concept originating from Greenshields resulted in the well-known speed-flow-density diagrams. Other references later in the literature are Daganzo (1997); Button (1993); Gazis (2002), etc.

In general, the study of this type of traffic flow models boils down to the analysis of these speed-flow-density relationships [see, e.g., the seminal paper of Greenshields (1935)]. The basic formula of traffic flow theory incorporates the interdependence of traffic flow q, traffic density k, and the speed v (Hall, 1996):

$$q = k \times v. \tag{2.1}$$

Using relation (2.1), the typical speed-flow-density diagrams can be constructed. Figure 1 shows the relationships between the speed-flow, the speed-density, and the flow-density diagram.

The actual form of these diagrams depends upon the prevailing traffic and roadway conditions on the roadway segment under study. Although the diagrams show continuous curves, it is unlikely that the full range of all points will be found at any particular measurement location. Almost all data collected for the calibration of these diagrams are subject to influences of changing environmental conditions, non-homogeneity of vehicles in the traffic stream, and lack of complete isolation from ramps and interchanges (Transportation Research Board, 1998). If traffic count data are available, traffic flows q can be assumed as given, which leaves us to calculate either traffic density or speed to complete the above formula. These diagrams also illustrate a number of significant points, i.e., a zero flow ($q = 0$) occurs under two very different conditions:

(1) When there are no cars in the facility, density is zero ($k = 0$), and flow is zero. Speed is purely theoretical for this condition and would be whatever the first

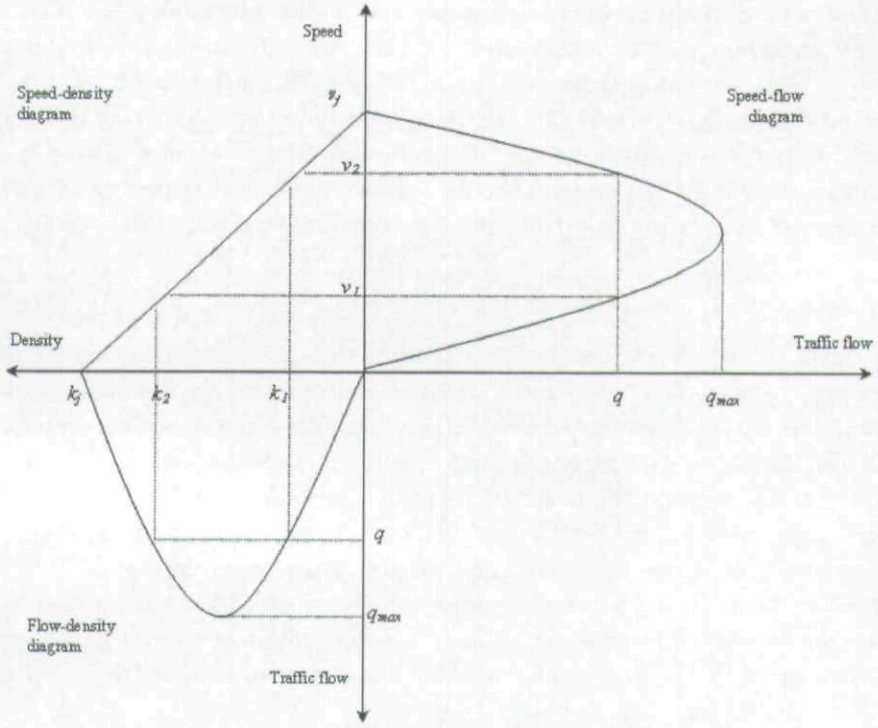

Fig. 1. The relations between the speed-flow, the speed-density, and the flow-density diagrams.

driver would select, probably the highest possible value (e.g., at the speed limit or v_f).

(2) When density becomes so high that all vehicles stop (speed is zero), the flow q is also zero. This is because there is no movement and vehicles cannot pass a point on the roadway. The density at which all movement stops is called the *jam density*, k_j.

Between these two extreme points, the dynamics of traffic flow produce a maximizing effect. As density k increases from zero, flow q also increases, since more vehicles are on the roadway. While this is happening, speed begins to decline (because of the interaction between vehicles). This decline is virtually negligible at low and medium densities and flows. As density continues to increase, these generalized diagrams suggest that speed decreases significantly before the capacity is achieved. Capacity is reached when the product of density and speed results in the maximum flow. Any flow other than capacity can occur under two conditions: one with a high speed and low density and one with low speed and high density. The high density and low speed side of the diagrams represents forced or breakdown flow (Daganzo, 1997). For each time period a different flow is observed and consequently different speeds and densities (following the speed-flow-density diagrams).

2.3. *Queueing theory approaches*

In the remainder of the paper, only traffic flow models based on queueing theory are presented and discussed. It is important to mention here that the queueing models reviewed here assume steady-state conditions, i.e., the same behavior is reproduced and observed every time with the same probability. In terms of traffic, this means that the traffic flows observed are stationary. The assumption of stationary traffic flows (or equivalent, steady-state queueing models) means that all vehicles will always be driving (no matter how slow). The non-stationary traffic however experiences stop's and go's (see, e.g., Boel and Mihaylova, 2006). These non-stationary traffic flows can be modeled using so-called transient queueing models. Heidemann (1999) shows that under non-stationary conditions, the speed-flow-density results deviate from the ones obtained with stationary queueing models. Moreover, phenomena like stop-and-go traffic can be explained using these models. Heidemann (1999) shows that the non-stationary flow-density diagrams converge to the stationary ones when the time period considered in the non-stationary models grows to infinity. In general, the steady-state results are most appropriate in design and policy recommendations. The transient queueing models are more useful in specific control situations for relatively small networks (Van Woensel, 2003).

Several papers are available in the academic literature that model traffic flows using a queueing approach: Heidemann (1996) showed that traffic flows could be adequately modeled using basic queueing models. Vandaele *et al.* (2000) elaborated further on Heidemann and extended his approach to general queueing models. In Heidemann (1996) and Vandaele *et al.* (2000), a basic framework for modeling traffic flows with queueing theory was developed. The queueing approach to traffic flows was explained in detail and applied to several single node (i.e., single stage) queueing models. Moreover, Van Woensel and Vandaele (2006) proved the validity of the queueing approach to uninterrupted traffic flows by comparing the queueing results with observed data on speed and flow. Van Woensel *et al.* (2006) came to the same conclusion based on the comparison of the queueing results with a traffic simulation study. In this paper, the relationship between flow and speed is simulated in order to test the effectiveness of some queueing based traffic models [see also Wuyts *et al.* (2004)]. These publications thus demonstrated the usability of the queueing models to adequately model traffic flows by comparing the queueing results with empirical data of speed and flow observed on highways and with simulation studies.

Jain and MacGregor Smith (1997), following a slightly different approach than the one in the previously mentioned papers, introduced the concept of state-dependent queueing models for modeling traffic flows and showed that it is more accurate in highly congested situations. In another paper, Cruz *et al.* (2005) stated that one of the most universal and significant applications where queueing networks occur include vehicular traffic flows. Finally, some research is done on a travel time-flow model originating from Davidson (1978). The model is based on some concepts of queueing theory but a direct derivation has not been clearly demonstrated

(Akçelik, 1991, 1996). Please note that there are many other publications that use queueing models but these are usually focused on other situations: interrupted traffic flows (Heidemann, 1991, 1994; Heidemann and Wegmann, 1997), where traffic is modeled on intersections with or without traffic lights; incident management (Qin and Smith, 2001; May, 1990), where queueing analysis is used to estimate traffic characteristics under incident situations, including the estimation of the maximum queue length, average queue length, maximum individual delay, average individual delay, and total delay or use oversimplified queueing models [e.g., deterministic queues in May and Keller (1957)].

In the queueing approach to traffic flow analysis, roads are subdivided into segments, with length equal to the minimal space needed by one vehicle on that road (Fig. 2). Define k_j as the maximum traffic density (i.e., maximum number of cars on a road segment). This length is then equal to $1/k_j$ and matches the minimal space needed by one vehicle on that road. Each road segment is then considered as a service station, in which vehicles arrive at a certain rate λ and get served at another rate μ (Vandaele *et al.*, 2000).

Vandaele *et al.* (2000) and Heidemann (1996) showed that the speed v can be calculated by dividing the length of the road segment $\frac{1}{k_j}$ by the total time in the system W.

$$v = \frac{1/k_j}{W}. \tag{2.2}$$

The total time in the system W is then different depending upon the queueing model used. The total time in the system W is then the sum of the waiting time W_q and the service time W_p, or:

$$\begin{aligned} W &= W_q + W_p \\ &= W_q + \frac{1}{k_j v_f}. \end{aligned}$$

In general, formula (2.2) can be rewritten in the following general form:

$$v = \frac{v_f}{1+\Omega}. \tag{2.3}$$

Formula (2.3) shows that the speed is only equal to the maximum speed v_f if the factor Ω is zero. For positive values of Ω, v_f is divided by a number strictly larger than 1 and speed is reduced. The factor Ω is thus the influence of congestion on speed. High congestion (reflected in a high Ω) leads to lower speeds than the

Fig. 2. Queueing representation of traffic flows.

maximum. The factor Ω is a function of a number of parameters depending upon the queueing model chosen: the traffic intensity ρ, the coefficient of variation of service times c_s and coefficient of variation of inter-arrival times c_a. High coefficients of variation or a high traffic intensity will lead to a value of Ω strictly larger than zero. Actions to increase speed (or decrease travel time) should then be focussed on decreasing the variability or on influencing the traffic intensity, for example by manipulating the arrivals (arrival management and ramp metering). The above function (2.3) is an exact analytical representation of the speed-flow-density diagrams. The specific shape of the speed-flow-density diagrams will depend upon the queueing model chosen and on the parameters used for the model (Vandaele *et al.*, 2000). Table 1 shows the specific form of Ω for the general queueing models.

The advantage of the queueing models is that the physical characteristics of the road network immediately can be mapped onto the parameters of the queueing model. The flow q is a parameter that is determined empirically over time, allowing the determination of realistic velocity profiles as a function of time. In the queueing formulas, one has four parameters that allows one to model any possible situation: the coefficient of variation of the inter-arrival times c_a, the coefficient of variation of the service times c_s, the jam density k_j and the free flow speed v_f. In practice, the jam density and the free flow speed are fixed for a given arc (i, j), leaving the coefficients of variation to represent the specific traffic conditions (e.g., bad weather, etc.), see Van Woensel and Vandaele (2006); Van Woensel *et al.* (2006) for some insights on how to determine these values in real-life.

2.4. *Added value and limitations of the queueing approach*

Queueing theory might be perceived as being too mathematically restrictive to be able to model all real-world situations. This handicap arises because the underlying queueing theory assumptions do not always hold in the real world. Any model (e.g., econometrical models, simulation, etc.) is an abstraction of reality. A queueing

Table 1. The specific form of Ω for each queueing model.

Queueing Model	Ω
$GI/G/1KLB$	$\frac{\rho}{(1-\rho)} \frac{(c_a^2+c_s^2)}{2} \frac{1}{k_j v_f} \exp\left[\frac{-2(1-\rho)(1-c_a^2)^2}{3\rho(c_a^2+c_s^2)}\right]$
$GI/G/zK$	$\left(\frac{c_a^2+c_s^2}{2}\right)\left(\frac{\rho^{\left(\sqrt{2(m+1)}-1\right)}}{m(1-\rho)}\right)\left(\frac{1}{k_j v_f}\right)$
$GI/G/zW$	$\phi\left(\frac{c_a^2+c_s^2}{2}\right)\left(\frac{1}{k_j v_f}\right) W_{qM/M/k}$
	With ϕ a correction factor defined in Whitt (1993) and $W_{qM/M/z}$ the formula for the waiting time in an $M/M/z$ queue.

model is a mathematical model and as a consequence, an abstraction of reality. In this review, it is shown however that this abstraction does not mean that the queueing approach is inadequate to model traffic flows and congestion.

On the other hand, depending upon the specific research question asked, the queueing approach might not be the best methodology. If one wants to explicitly take into account the complexity inherent in the traffic domain such as modeling travellers behavior and their interaction with Advanced Travel Information Systems, alternative modeling approaches are needed. This social nature of traffic operations seems to be a key question more and more explored in the literature (Erol *et al.*, 2005). Usually drivers are modeled as interacting social agents allowing their behavior to be predicted and considered in an agent-based micro-simulation approach (Adler *et al.*, 2005), where a complex system is viewed as a large set of small, interacting components. The main focus is on identifying the components in a system, discovering their local behavior and their interactions. Two issues with agent-based micro-simulation are the computational performance, and the software development cost. Micro-simulations running at a very detailed level, emulating the individual behavior of every entity in the system are thus computationally very intensive (Erol *et al.*, 2005).

If one accepts the smaller level of behavioral insights, the queueing approach also offers some advantages when incorporated in decision and/or optimization tools: it provides both insight and explanation; it is comprehensive and because it involves explicit analytical modeling, it is very well suited for serving in higher level problems with optimization purposes. From a practical point of view and given the inherent possibility for optimization, the queueing approach is a tractable method if a traffic model has to be embedded in other models with optimization purposes, e.g., dynamic vehicle routing problems (VRP). The main difficulty in the VRP models, is that one needs a good representation of the traffic process itself that easily can be integrated. The key issue is thus the characterization of the speed distribution and the resulting travel times on each link over different time slices due to the stochastic and dynamic nature of travel times. Here, the queueing approach is extremely valuable [see, e.g., Van Woensel *et al.* (2007); Kerbache and Van Woensel (2005) for more details]. The queueing models developed also contribute enormously to the validation and evaluation of economic policies. A first attempt to combining economics and the queueing models is a congestion pricing application presented in Van Woensel *et al.* (2005).

3. Traffic Flows and Queueing Networks

In general, queueing networks are defined as open, closed, or mixed. In open queueing networks, customers enter the system from outside, receive some service at one or more nodes and then leave the system. In closed queueing networks, customers never leave or enter the system: a fixed number of customers circulate within the network (Whitt, 1984). Mixed queueing networks are systems that are open with

respect to some customers and are closed with respect to other customers (Balsamo *et al.*, 2001). Research in the area of queueing networks is very active, resulting in a vast amount of journal papers, books, reports, etc. For general and complete classifications of queueing networks, the reader is referred to, e.g., Walrand (1988). If one assumes that vehicles arrive from outside the network, follow a route in the network and finally, leave the system then open queueing networks are most appropriate to model road networks. As a consequence, closed and mixed queueing networks are not discussed in this paper. The interested reader is referred to the following references on closed and mixed queueing networks: (Perros, 1994); Suri *et al.* (1993); Buzacott and Shanthikumar (1993); Kleinrock (1975); Kelly (1979) and many others.

It is assumed that vehicles arrive in the network following a general distribution and are served according to a general distribution. Moreover, for traffic flows on highways, only feed-forward flows through the network need to be considered: i.e., the only way of arriving at a node is from an immediately preceding connected node or from outside the system. This means that loops or feedbacks in the road network are not considered. Of course, other distributions (than the general distribution) can be assumed too for the arrival and service rates. In this paper it is however opted to work at each node m with general distributions for the arrival and service rate with a mean rate of λ_m and μ_m respectively and a coefficient of variation of c_{Am} and c_{Sm} respectively. Summarizing, the road network can be modeled as a feed-forward open generally distributed queueing network with one (or more) type(s) of vehicle (multi-class) and individual arrivals of each vehicle.

Vandaele *et al.* (2000) proved that the average speed at node m denoted by v_m, can be obtained as (or adding a subscript m for the node to formula (2.2)):

$$v^m = \frac{1/k_j^m}{W^m}, \tag{3.1}$$

with k_j^m the jam density at node m and W^m the total time spent in node m. This total time is then the sum of both the waiting time at node m: W_q^m and the service time at node m: W_p^m. The problem now shifts to finding an expression for W^m (or equivalently for both W_q^m and W_p^m). The analysis is now split up into two parts depending upon the buffer size: infinite buffer sizes (Sec. 3.1) versus finite buffer sizes (Sec. 3.2).

3.1. *Infinite buffer sizes*

In this section, the assumption is made that capacity of the buffer space between two consecutive connected nodes is infinite. A buffer can then accommodate any number of customer waiting for service (Perros, 1994). As a consequence, each node in the network is unaffected by events occurring at other nodes. Therefore, the insights and formulas obtained for the single node models can be utilized [see Vandaele *et al.* (2000)]. The methodology for modeling infinite queueing networks and obtaining the relevant performance measures is presented in this section.

Unfortunately, for the $GI/G/1$ and $GI/G/z$ queueing models, no closed form expression for the expected waiting time in the system W_q^m is available. Consequently, to determine W_q^m for the $GI/G/1$ and $GI/G/z$ queueing models, one has to rely on approximations. These approximations make use of the mean and the variance of the inter-arrival and process time distributions. In the literature, numerous approximations are described: e.g., Kraemer and Lagenbach-Belz (1976); Marchal (1976); Page (1972); Kingman (1964) (also known as the heavy-traffic approximations), Whitt (1993), etc. The Kraemer-Lagenbach-Belz approximations (KLB) are generally considered as a good approximation (Buzacott and Shanthikumar, 1993) but are restricted to one server only. To cope with multiple servers, the Whitt approximations (W) are used. The Kingman approximations (K) are interesting because they perform well under the assumption of high traffic intensity which typically occurs at highly congested situations (i.e., the Kingman approximation is also referred to as a heavy-traffic approximation). For a discussion, comparison and further references of different approximations the reader is referred to Van Woensel (2003) and Vandaele (1996).

The performance measures of queueing networks depend upon the inter-arrival times and the service requirements, and the related coefficients of variation. It is supposed that the inter-arrival and service distributions of flows entering the network to be well specified. The problem now is to completely characterize the movements of vehicles between the different nodes. More specifically, insights are needed on how the service characteristics of one node affects the inter-arrival process at other nodes (Zijm, 2002). Before continuing, the following variables are defined:

c_{Am} = Coefficient of variation of inter-arrival time at node m;
λ_t = Aggregated arrival rate at node t;
ρ_t = Utilization ratio at node t;
c_{At} = Coefficient of variation of inter-arrival time at node t;
c_{St} = Coefficient of variation of service time at node t;
c_{Dt} = Coefficient of variation of departure time at node t;
λ'_m = Arrival rate from flow outside the system arriving at node m;
c'_{Am} = Coefficient of variation of inter-arrival time from flow outside the system arriving at node m;
λ_m = Arrival rate at node m from flows within the system;
$p_{t,m}$ = The proportion of the flow arrived in node t going to node m;
$c_{At,m}$ = Coefficient of variation of inter-arrival time at node t going to node m.

In networks, the coefficient of variation of the inter-arrival times at node m depends on the departure processes of all preceding nodes. Of course, these departure processes are a function of both the respective arrival and service processes at these preceding nodes. Assume m is the node under consideration and all nodes preceding this node m are indexed with the letter t. Due to the dependency on preceding nodes, an equation which relates the coefficient of variation of the inter-arrival times at node m with the properties of the preceding machines is needed. The following

relationship captures this concern (Vandaele, 1996):

$$c_{Am}^2 = \sum_{t=1}^{T} \left(\frac{\lambda_t}{\lambda_m} p_{t,m} \right) c_{At,m}^2 + \frac{\lambda'_m}{\lambda_m} c_{Am}'^2. \tag{3.2}$$

The first term of expression (3.2) relates to a weighted average of the aggregate squared coefficient of variation of the internal arrivals to node, the second term takes into account the aggregate squared coefficient of variation of the external arrivals at node m. In other words, the coefficient of variation of the inter-arrival times is the sum of the weighted average of the coefficient of variation of the inter-arrival times coming from all preceding nodes (connected with node m) and the weighted average of the coefficient of variation of the inter-arrival times coming from outside the system. The coefficient of variation of the inter-arrival times coming from all preceding nodes (connected with node m) is a function of the departure processes of these nodes. The second term in expression (3.2) takes into account the squared coefficient of variation of the streams coming from outside the system. The aggregated arrival rate at node $m(\lambda_m)$ is defined as the sum of all arrival rates entering node m:

$$\lambda_m = \sum_{t=1}^{T} \lambda_t p_{t,m} + \lambda'_m.$$

Buzacott and Shanthikumar (1993) report the following relationship between $c_{At,m}$ and c_{Dt}:

$$c_{At,m}^2 = p_{t,m} c_{Dt}^2 + (1 - p_{t,m}). \tag{3.3}$$

If all traffic goes from node t to node m, the proportion $p_{t,m}$ equals 1. As a consequence, relationship (3.3) reduces to: $c_{At,m}^2 = c_{Dt}^2$. If however, $p_{t,m} = 0$ (or there is no direct flow between node t and node m) relationship (3.3) reduces to $c_{At,m}^2 = 1$.

On its turn, the squared coefficient of variation of departure time at node $t(c_{Dt}^2)$ is approximated using the following expression (Hopp and Spearman, 1996):

$$c_{Dt}^2 \approx \left(1 - \rho_t^2\right) c_{At}^2 + \rho_t^2 c_{St}^2. \tag{3.4}$$

Formula (3.4) shows that the departure process is a function of both the arrival process and the service process at node t. When traffic intensity is high (ρ_t^2 is high) at node t, the service process will have a larger impact on the departure process. In the extreme case when $\rho_t^2 = 1$, $c_{Dt}^2 \approx c_{St}^2$. The opposite is true when having low traffic intensity (ρ_t^2 is low): the arrival process then determines the departure process at node t. In the extreme case when $\rho_t^2 = 0$, $c_{Dt}^2 \approx c_{At}^2$. This expression originates from Kuehn (1979) which is originally based on the Marshall (1968) formula.

Substituting expressions (3.3) and (3.4) in formula (3.2), some minor reworking and expressed as an equality, the following formula for the squared coefficient of

variation of the inter-arrival times at node m can be approximated by the following formula:

$$c_{Am}^2 = \frac{1}{\lambda_m} \left\{ \sum_{t=1}^{T} \left[\begin{array}{l} \lambda_t p_{t,m}^2 \left(1 - \rho_t^2\right) c_{At}^2 + \lambda_t p_{t,m} \\ \quad \times \left(p_{t,m} \rho_t^2 c_{St}^2 + 1 - p_{t,m}\right) + \lambda'_m c'^2_{Am} \end{array} \right] \right\}. \tag{3.5}$$

There are as many of these equations as there are nodes in the queueing/traffic network. In order to find the coefficients, this set of linear equations has to be solved simultaneously. However, due to the assumption of feed-forward in flow (without loops and feed-backs) a single pass methodology can be used. As the only information needed to solve Eq. (3.5) are the variables of the immediate preceding nodes, these equations can be solved in a straightforward way: i.e., start at the first node and work your way through the network to the last node.

3.2. *Finite buffer sizes*

In this section, the step to queueing networks with finite buffer sizes is made: the assumption is now that the capacity of the buffer space between two consecutive connected service stations is finite. As a consequence, each node in the network might be affected by events at other nodes, leading to the phenomena of blocking and starvation (Perros, 1994).

In general, two blocking mechanisms can be distinguished: blocking-after-service and blocking-before-service. Blocking-after-service occurs when after service, a customer sees that the buffer in front of him is full and as a consequence cannot continue its way in the network. Blocking-before-service implies that a server can start processing a next customer only if there is a space available in the downstream buffer. If not, the customer has to wait until a space becomes available. Most production lines operate under the blocking-after-service system. Moreover, in the literature it is the most commonly made assumption regarding the buffer behavior (Dallery and Gershwin, 1992). For the road networks discussed in this paper, the blocking-after-service mechanism is assumed.

In general, four possible approximations methodologies for solving finite queueing network models can be used (Jain and Macgregor Smith, 1994): isolation methods, repeated trial, node-by-node decomposition and expansion methods. In this paper, the expansion method is used and involves three stages: network re-configuration, parameter estimation and feedback elimination. The expansion method is developed to solve $M/M/z/K$, $M/G/z/K$ and $GI/G/1/K$ queueing networks. Unfortunately, no papers exist yet in the literature that report on approximations for $GI/G/z/K$ queueing networks. Consequently, when using the expansion method with general inter-arrival distributions and general service distributions all lanes on the road are aggregated into one lane.

Instead of presenting a formal overview of the queueing methodology, this section will offer an intuitive approach and a selection of formulas that is sufficient to understand the remainder of this paper. Moreover, the flexibility of the expansion

method allows for easy adding newly developed formulae. In the remainder of this section, the expansion method will briefly be described. For more information and applications of the (generalized) expansion method, the reader is referred to Kerbache and Macgregor Smith (1987, 1988); Cheah and MacGregor Smith (1994); MacGregor Smith and Cruz (2005a, 2005b); Spinellis *et al.* (2000); Cruz and MacGregor Smith (2007).

For finite queueing networks it can be proven by Little's adapted formula that the total time in the node W^m is obtained by the following expression:

$$W^m = \frac{L^m}{\lambda_m(1-p_K^m)}, \tag{3.6}$$

with L^m the average number of customers in the node and p_K^m blocking probability at node m with a buffer size K at node m. Instead of following Kerbache and MacGregor Smith (2000), the blocking probability p_K^m is now obtained as follows (Kim and Chae, 2003):

$$p_K^m = (1-\rho_m)\left[\left(\frac{a_R}{a_R+b-a}\right)\left(1+\frac{a-b}{a_R+b_R-a}\right)^{K-1} - \rho_m\right]^{-1}. \tag{3.7}$$

The number of customers L^m is obtained using the following formula (Kim and Chae, 2003):

$$\begin{aligned} L^m = {} & \frac{\lambda_m(a_R+b_R-a)}{1-\rho_m} + \frac{b-b_R}{a_R} \\ & + \left[\frac{\rho_m(b-b_R)}{(1-\rho_m)a_R} - \frac{\lambda(a_R+b_R-a)+\rho_m K}{1-\rho_m}\right] p_K^m \\ & + (1-p_K^m)(1-\lambda_m a_R) \end{aligned} \tag{3.8}$$

with $a = \frac{1}{\lambda_m}$, $b = \frac{1}{\mu_m}$, $a_R = \frac{(c_{Am}^2+1)}{2\lambda_m}$ and $b_R = \frac{(c_{Sm}^2+1)}{2\mu_m}$ (Kim and Chae, 2003). The difficulty is now to obtain the values of the unknowns: c_{Am}^2, ρ_m. These values can be determined by solving the finite queueing network using the expansion method (Kerbache and Macgregor Smith, 1987).

The first step in the expansion method involves re-configuring the network: an artificial node is added for each finite node in the network. The artificial node is added to register the blocked customers at the finite node. In Fig. 3, node 2 has a finite capacity, i.e., the buffer space between node 1 and 2 is limited to a fixed value K. Therefore, a holding node h is included to account for the blocked customers (Kerbache and Macgregor Smith, 1988).

If node 2 is not saturated (i.e., the buffer is not full), the customer proceeds to the queue of node 2, with probability $1-p_K$ (with p_K the probability that there are K customers waiting in the queue or the queue is full). However, if node 2 is saturated (i.e., the buffer is full) then the customer is blocked with probability p_K and is redirected to the artificial node. As a consequence, this customer will incur a delay before he can join the queue of node 2. After this delay, the customers again tries to

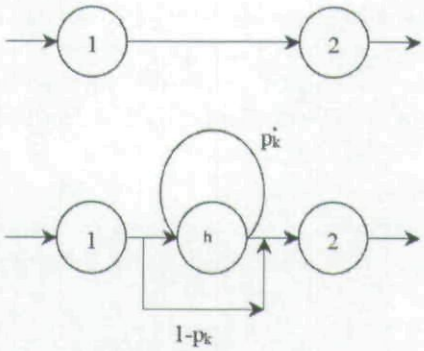

Fig. 3. The expansion of a finite queue.

join the queue of node 2. Now two possible situations can occur: either the queue is still full and the customer incurs another delay with probability p'_K, or the customer enters the queue to receive service at node 2. To represent this process, the artificial node has a feedback arc to account for these attempts. The artificial node has the same arrival and service process as the finite node for which it is inserted. The only difference is that there are ∞ servers, so that a blocked customer immediately can receive its service/delay (Kerbache and Macgregor Smith, 1988).

In the second step, the parameters for the system described above are determined. These unknowns are in the most general case (Kerbache and Macgregor Smith, 2000): the probability of a customer being blocked: p_K, the probability that a customer is forced back to the holding node given he was rejected at the previous trial: $p_K^{'}$, the service rate at the holding node: μ_h, the squared coefficient of variation of the departure process from node 1: c^2_{D1}, the squared coefficient of variation of the arrival process to the blocked node 2: $c^2_{a1,2}$, the squared coefficient of variation of the arrival process at the holding node h: c^2_{ah} and the squared coefficient of variation of the service process at the holding node h: c^2_{sh}. These parameters can be obtained by the formulae described in Kerbache and MacGregor Smith (2000) and the references mentioned therein.

The repeated visits to the holding nodes (due to the feedbacks), create strong dependence in the arrival process. Therefore, the repeated immediate feedback needs to be eliminated. This is done by giving the customer a total service time during the first passage through the holding node. This last step is called the feedback elimination in Kerbache and MacGregor Smith (2000).

Summarizing, the expansion method involves first expanding the network by adding a holding node for each finite node, then setting up a system of nonlinear equations combining the above relationships and finally, removing the feedback at all holding nodes (Kerbache and Macgregor Smith, 1987). The system of nonlinear equations is dependent upon the specific assumptions regarding the distribution of the arrival and service process. For $GI/G/1/K$ queueing networks a system of nonlinear equations needs to be solved with the following unknowns: $\{\lambda\ \lambda_j\ \lambda_h\ p_K$

$p'_K\ c^2_{di}\ c^2_{aih}\ c^2_{aj}\ c^2_{sh}\}$ (Kerbache and Macgregor Smith, 1987, 2000). This system of nonlinear equations for the $GI/G/1/K$ queueing network is programmed in Wolfram Mathematica 5.0 and can be solved using a Newton-like nonlinear optimization technique [see, e.g., Malek-Madani (1997)].

4. Demonstration of the Methodology

In the previous section, different formulas for modeling uninterrupted traffic flows using queueing networks with infinite buffer sizes and finite buffer sizes were developed. In this section, these formulas will be used for some basic network layouts: first, the basic layouts will be discussed and secondly, an arbitrarily configured network is analyzed.

4.1. *Basic layouts*

The following three network sample layouts are considered: Tandem network configurations, Merge network configurations and Split network configurations. Any possible network is always a combination of these three basic network layouts. Therefore, it is interesting to study them separately. Figure 4 presents the network configurations in more detail.

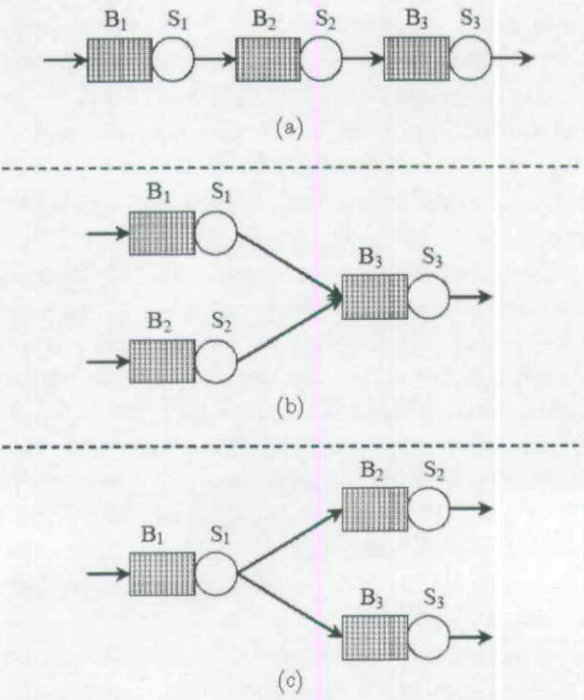

Fig. 4. The different network layouts.

Table 2. The parameters used for the nodes in the networks.

	k_j	v_f	c_a	c_s	K
Setting 1	120	120	0.7	0.5	$+\infty$
Setting 2	120	120	0.7	0.5	360
Setting 3	120	120	0.7	0.5	240
Setting 4	120	120	0.7	0.5	120
Setting 5	120	120	0.7	0.5	60

The formulas for the queueing networks with infinite buffer sizes are programmed in Microsoft Visual Basic 6.0 and can be manipulated using a graphical user interface. All formulas for the queueing networks with finite buffer sizes are programmed in Mathematica 4.0. This software tool is chosen because the solution of these models requires nonlinear optimization techniques readily available in Mathematica 4.0. Each network layout and its results is discussed in detail.

Table 2 presents the input data used for different parameter settings depending upon the buffer size. These values come from an empirical analysis where observed data is compared to the queueing model results (Van Woensel and Vandaele, 2005). The buffer size numbers are obtained as follows: focusing on setting 1, it is assumed that each road segment has an equal length of 3 km: i.e., the space between server i and server $i+1$ is 3 km, or the length of the buffer before server $i+1$ is equal to 3 km. As a consequence, the buffer capacity is equal to the buffer length, 3 km multiplied with the maximum traffic density k_j, 120 veh/km, resulting in 360 vehicles. In this calculation, it is assumed that all lanes are aggregated into one lane [similar to the literature, e.g., Daganzo (1997)]. The other parameter settings follow the same logic for the calculation of the buffer size, but the length of the segment is shortened from 3 km to 0.5 km (in setting 5). It should be noted that the purpose of this demonstration is to show the effect of the buffer size on the queueing results, rather than presenting a complete experimental design on all parameters used in this analysis. For the latter, the reader is referred to Van Woensel (2003).

Each parameter setting is applied to each of the above network layouts. The buffer capacity (K_i) for network models with infinite buffers is set to infinity ($K_i = +\infty$, for each node i). However, for the network models with finite buffers, the values K_i in the table are used for the capacity. Also the parameter c_a is only used when applicable, i.e., an arrival of vehicles from outside the network. The arrival of vehicles from outside the network is for layout 1 equal to 4,500 vehicles at node 1; for layout 2, 2,500 vehicles at both nodes 1 and 2 and again 4,500 vehicles at node 1 in layout 3. The routing probability from nodes 1 and 2 connecting 3 in network layout 3 is 0.5. The results for the above network layouts combined with the different scenarios are presented in Table 3.

As one can see, the results between setting 1 and setting 2 do not differ much. The reason for this observation is that the blocking probability in setting 2 is very small. As a consequence, the difference with the results for the infinite buffers network are negligible. When the buffer between two connected nodes gets smaller,

Table 3. Results for the different settings and network layouts.

			Setting 1	Setting 2	Setting 3	Setting 4	Setting 5
Layout 1	(B_1, S_1)	c_a	0.7	0.7	0.7	0.7	0.7
		v	109	109	107	100	96
		k	41	41	42	45	47
	(B_2, S_2)	c_a	0.69	0.69	0.72	0.75	0.80
		v	111	111	105	97	89
		k	41	41	43	46	51
	(B_3, S_3)	c_a	0.67	0.67	0.73	0.76	0.83
		v	112	112	104	95	87
		k	40	40	43	47	52
Layout 2	(B_1, S_1)	c_a	0.7	0.7	0.7	0.7	0.7
		v	119	119	118	113	105
		k	21	21	21	22	24
	(B_2, S_2)	c_a	0.7	0.7	0.7	0.7	0.7
		v	119	119	118	113	105
		k	21	21	21	22	24
	(B_3, S_3)	c_a	0.66	0.66	0.74	0.79	0.86
		v	109	109	103	101	94
		k	46	46	49	50	53
Layout 3	(B_1, S_1)	c_a	0.7	0.7	0.7	0.7	0.7
		v	109	109	107	100	96
		k	41	41	42	45	47
	(B_2, S_2)	c_a	0.83	0.83	0.85	0.91	0.98
		v	118	118	116	113	109
		k	19	19	19	20	21
	(B_3, S_3)	c_a	0.83	0.83	0.85	0.91	0.98
		v	118	118	116	113	109
		k	19	19	19	20	21

the difference becomes more significant: the blocking probability gets larger, resulting in a decrease in speeds. More specifically, decreasing buffer sizes, results in lower speeds.

When focusing on layout 1 in setting 1, one can see that if no splits, merges, off-ramps, or on-ramps are present and all nodes are identical, speed increases and density decreases. It seems that flow goes to a certain structured stable state if there are no interruptions. The driving factor behind this phenomenon is the decreasing coefficient of variation of inter-arrival times. Uncertainty decreases and, as a consequence, speed increases and density decreases. It can be shown (Van Woensel, 2003) that if the tandem of nodes would be very long, the coefficient of variation of inter-arrival times will approximate the coefficient of variation of service times, leading to a stable speed and density. In this case, more than 20 identical nodes in the network leads to a coefficient of variation of inter-arrival times equal to 0.5. Speed will be almost 120 kilometer per hour and density would be 37 vehicles per kilometer. In the other settings studied, this effect is disturbed by the corrupting influence of the blocking probability.

When looking at the merge layout in setting 1, some observations are made. At the merge node, the coefficient of variation of the inter-arrival times decreases very

little. This is mainly due to the small coefficients of variation of the service times. This decrease in the coefficient of variation is however offset by the merging of the flow (5,000 vehicles), resulting in a lower speed (109 km/hr) than the individual merging streams. In the other settings studied for the merge layout, this effect is even increased by the influence of the increasing blocking probability.

Finally, based on the split layout for setting 1, important insights can be reported. After the split node, the coefficient of variation of the inter-arrival times to the connected nodes increases. The increase in the coefficient of variation is however offset by the splitting of the flow (2,500 vehicles), resulting in a higher speed (118 km/hr) than the stream before the split. In the other settings studied for the split layout, this effect is again increased by the influence of the increasing blocking probability.

4.2. *Arbitrarily configured networks*

An arbitrarily configured network is presented as a last demonstration of the methodology (Fig. 5). This network can be seen as a combination of tandem, split and merge configurations. The nodes are defined in Table 4. The flow proportions are $p_{1,3} = 0.4$, $p_{1,4} = 0.6$, $p_{2,5} = 1.0$, $p_{4,6} = 1.0$, $p_{5,6} = 1.0$.

The buffer capacity (K_i) for network models with infinite buffers is set to infinity ($K_i = +\infty$, for each node i). Each road segment is assumed to have length of 3 km: i.e., the space between server i and server $i+1$ is 3 km, or the length of the buffer before server $i+1$ is equal to 3 km. For example, the buffer capacity at node 1 is equal to the buffer length, 3 km multiplied with the maximum traffic density k_j, 90 veh/km, resulting in 270 vehicles. For the network models with finite buffer sizes this results in the following values $(K_1, K_2, K_3, K_4, K_5, K_6) = (270, 180, 150, 210, 210, 210)$ are used for the capacity of these buffers. After a computer run of a few seconds in Microsoft Visual Basic, the output in Table 4 is generated. The probability of blocking is very small. As a consequence, the difference with the results for the infinite buffers network are negligible. Experiments in Van Woensel (2003) show the effects of the blocking parameter on the speeds. Basically, the conclusions made in the previous section hold.

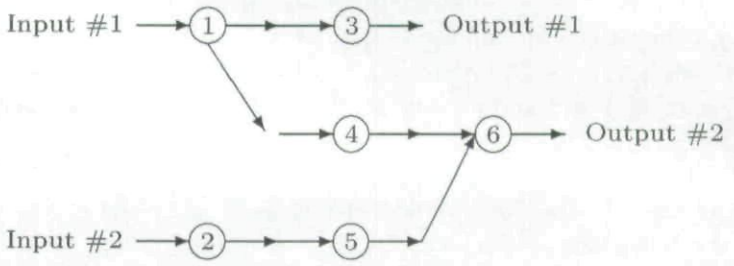

Fig. 5. An arbitrarily configured network.

Table 4. The arbitrarily configured network: Input and results.

Node	Input					Infinite Bufers	Finite Bufers	
	c_a	c_s	k_j	v_f	Flow	Speed	p_K	Speed
1	0.50	0.5	90	120	3000	120	4.01×10^{-21}	120
2	0.40	0.5	60	120	1550	120	2.85×10^{-21}	119
3	0.86	0.5	50	120	1200	112	7.53×10^{-20}	112
4	0.81	0.5	70	120	1800	112	3.23×10^{-21}	111
5	0.64	0.5	70	120	1550	120	3.04×10^{-21}	120
6	0.68	0.5	70	120	3350	75	3.12×10^{-21}	75

5. Optimal Spacing of Counting Points on the Road

More and more detector loops are being installed throughout the world to count the number of vehicles passing a certain point (Boel and Mihaylova, 2006). Mostly, the decision concerning the location of the detector is arbitrarily (e.g., near an off-ramp or on-ramp). For traffic management purposes these detector loops should be placed on strategic places to assure the best follow-up and traffic guidance. Based on the insights obtained from the literature on finite versus infinite queueing networks applied to traffic environments, a policy tool is developed to determine the optimal positions of the vehicle detectors on highways. In other words, what are the optimal positions of the counting points on congested highways?

Formally, the question can be reformulated as follows: what should be the buffer capacity such that the blocking probability is equal to some threshold value. This threshold value for the blocking probability should be set in such a way that an influence on the performance measures (and consequently, speed) can be derived. In this section, it is assumed that the threshold value for the blocking probability (p_K^t) is equal to 0.005. Of course, any other value could be chosen.

$$\text{Max } K$$

$$\text{subject to:}$$

$$p_K \geq p_K^t.$$

The threshold value for the blocking probability p_K^t is set to 0.005 in the analysis. The specific results for the buffer size K now depend upon the queueing model and its specific parameters chosen. In a second step, this buffer size K should be transformed in a measurement in meters to obtain the spacing. This is done by using the maximum traffic density k_j. The buffer size K can be calculated as $K = k_j \times \text{length} \times \text{lanes}$. Assuming all lanes are aggregated into one lane (i.e., lanes = 1) and minor reworking, the length of the segment is equal to: $\text{length} = \frac{K}{k_j}$.

The maximum traffic density k_j is then obtained from the specific queueing model used. Figure 6 shows for different values of the traffic intensity ρ, the maximum corresponding space needed for the counting points to have a blocking probability of at least 0.005.

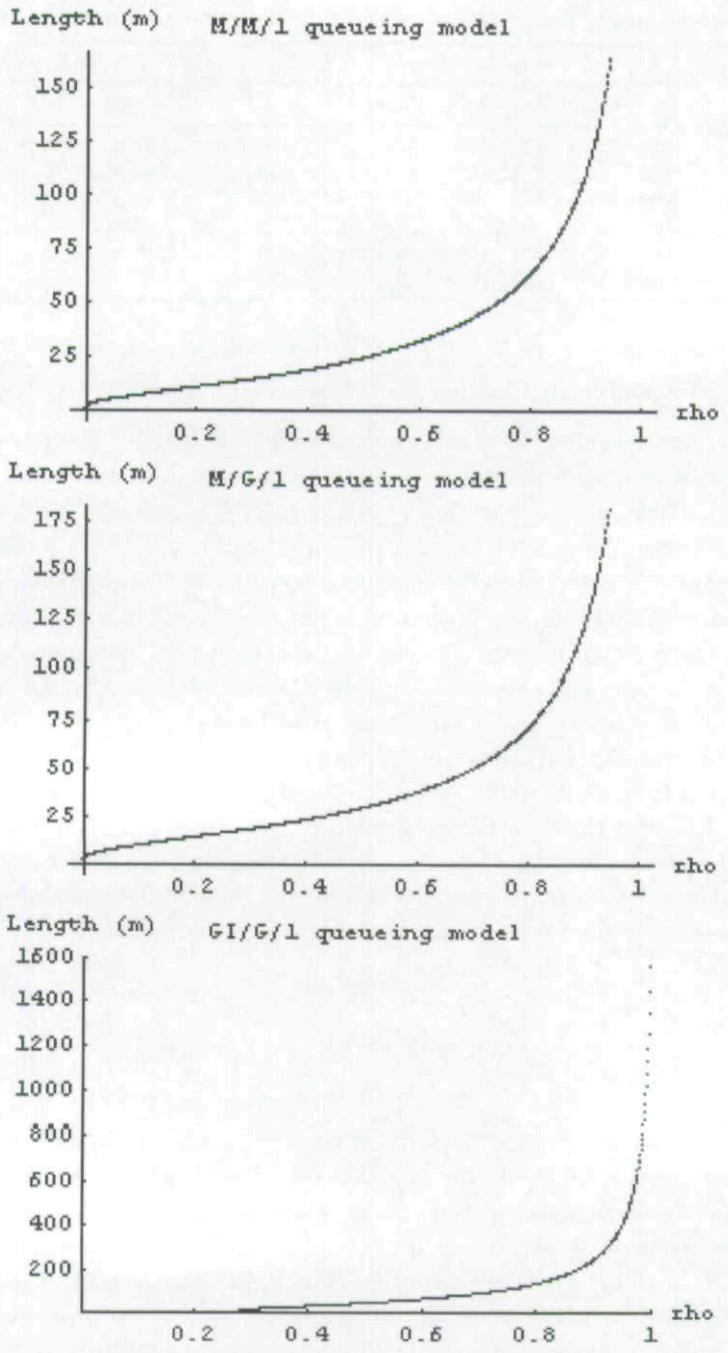

Fig. 6. Overview of the optimal spacing as a function of the traffic intensity ρ for different queueing models.

For low traffic intensities ($0 \leq \rho \leq 0.6$), the infinite buffer models would be a better option to use. However for larger traffic intensities ($\rho > 0.6$), the functions in Fig. 6 can be used to determine the maximum space between two consecutive counting points. For example, assuming a traffic intensity of 0.8 and the parameters set to the value in the above table, the optimal spacing would be every 146 meter. Of course, in highly congested situations the variability will be higher than the value of 0.75 chosen above. Assuming that the coefficients of variation of the inter-arrival times and the service times are both equal to 1.5, the result changes to 572 meter for the $GI/G/1$ queueing model. The latter value is close to spacing (approximately 500 meter) used between the counting points on a highly congested highway ($E17$ Gent to Antwerp, segment 8 km before the Kennedy tunnel) in Belgium (De Schutter *et al.*, 1999).

The threshold value was set to 0.005 in the above analysis. Experiments show that increasing this value (i.e., pursuing a higher blocking probability) leads to a closer-to-each-other spacing. In the same way, a smaller threshold for the blocking probability, leads to larger segments between the counting points.

6. Future Research Opportunities

An interesting future research step could be the empirical validation of the queueing networks. Unfortunately, no data for a complete road network (including all links, on-ramps and off-ramps) was available at this moment. As a consequence, no validation of the network models could be done directly. In previous work (Van Woensel and Vandaele, 2006), it was demonstrated that the queueing models for single nodes are very good approximations for the real traffic flows. Because the network models with infinite buffer sizes consist of the formulas for the single nodes, it is expected that the validation of these network models will be positive. An interesting extension would then be to evaluate the performance of the finite queueing networks.

An interesting approach is the embedding of the queueing approach in the agent-based simulation environment. Cetin *et al.* (2003) use a simple queueing model to limit the number of agents that can be on a link at the same time. It would be worthwhile to incorporate the more advanced queueing models described in this review in the agent-based models. As such, one could get the best of two worlds: the number of agents on a link is determined by queueing models, while the interaction between the agents comes from the simulation.

7. Conclusions

In this review, the queueing approach to traffic flows is discussed in great detail. Starting from the general insights from traffic flow theory, the link with queueing models is made. Most of these queueing models published so far in the literature are based on single node analysis. An extension towards queueing networks based on the well-available queueing literature is presented. The analysis was split up in two

parts depending upon the buffer size. Both the methodology for modeling infinite and finite buffers queueing networks was explained in detail. Results showed that if the buffer size is sufficiently large enough, results for finite and infinite buffers networks are almost the same.

This paper ended with an simple application of the finite queueing network methodology. A tool was presented to facilitate the public policy maker to determine the optimal positioning of the vehicle detectors. The main purpose was to determine the optimal spacing of the counting stations on a highway in order to get more detailed and useful information from the observed data in congested situations.

Acknowledgments

The authors would like to thank the two anonymous referees and the editor for their valuable input during the process of (re-)writing the paper.

References

Adler, JA, G Satapathy, V Manikonda, B Bowles and VJ Blue (2005). A multi-agent approach to cooperative traffic management and route guidance. *Transportation Research Part B*, 39, 297–318.

Akçelik, R (1991). Travel time functions for transport planning purposes: Davidson's function, its time-dependent form and an alternative travel time function. *Australian Road Research*, 21(3), 49–59.

Akçelik, R (1996). Relating flow, density, speed and travel time models for uninterrupted and interrupted traffic. *Traffic Engineering and Control*, 37(9), 511–516.

Balsamo, S, V de Nitto Personé and R Onvural (2001). *Analysis of Queueing Networks with Blocking*. Kluwer Academic Publishers.

Ben-Akiva, M, M Bierlaire, H Koutsopoulos and R Mishalani (1998). DynaMIT: A simulation-based system for traffic prediction. In: *Proceedings of the DACCORD Short-Term Forecasting Workshop*, Delft University.

Boel, R and L Mihaylova (2006). A compositional stochastic model for real time freeway traffic simulation. *Transportation Research Part B*, 40, 319–334.

Branston, D (1976). Models of single lane time headway distributions. *Transportation Science*, 10, 125–148.

Button, KJ (1993). *Transport Economics*. Edgar Elgar Publishing.

Buzacott, J and JG Shanthikumar (1993). *Stochastic Models of Manufacturing Systems*. Prentice-Hall.

Cetin, N, A Burri and K Nagel (2003). A large scale agent-based traffic microsimulation based on a queue model. STRC, Third Swiss Transport Research Conference, March 19–21, 2003

Cheah, J and J MacGregor Smith (1994). Generalized M/G/C/C state dependent queueing models and pedestrian traffic flows. *Queueing Systems and Their Applications (QUESTA)*, 15, 365–386.

Cruz, FRB, J MacGregor Smith and RO Medeiros (2005). An M/G/C/C state-dependent network simulation model. *Computers & Operations Research*, 32, 919–941.

Cruz, FRB and J MacGregor Smith (2007). Approximate analysis of M/G/C/C state-dependent queueing networks. To appear in *Computers & Operations Research*.

Daganzo, CF (1997). *Fundamentals of Transportation and Traffic Operations*. Elsevier Science.

Dallery, Y and SB Gershwin (1992). Manufacturing flow line systems: A review of models and analytical results. *Queueing Systems*, 12, 3–94.

Davidson, KB (1978). The theoretical basis of a flow-travel time relationship for use in transportation planning. *Australian Road Research*, 8(1), 32–35.

De Palma, A and F Marchal (2002). Real cases applications of the fully dynamic METROPOLIS tool-box: An advocacy for large-scale mesoscopic transportation systems. *Networks and Spatial Economics*, 2, 347–369.

De Schutter, B, T Bellemans, B De Moor, S Logghe, J Stada and B Immers (1999). Advanced traffic control on highways. *Journal A*, 40(4), 42–51.

Ehlert, A, MGH Bell and S Grosso (2005). The optimisation of traffic count locations in road networks. To appear in *Transportation Research Part B*.

Erol, K, R Levy and J Wentworth (2005) Application of agent technology to traffic simulation. Highway Research Center, Available at: http://www.tfhrc.gov/advanc/agent.htm.

Gazis, DC (2002). *Traffic Theory*. Kluwer Academic Publishers.

Greenshields, BD (1935). A study of traffic capacity. *Highway Research Board Proceedings*, 14, 448–477.

Hall, F (1996). *Traffic Flow Theory: A State-of-the-Art Report*, Chapter Traffic Stream Characteristics. Transportation Research Board.

Heidemann, D (1991). Queue length and waiting-time distributions at priority intersections. *Transportation Research Part B*, 25, 163–174.

Heidemann, D (1994). Queue length and delay distributions at traffic signals. *Transportation Research Part B*, 28, 377–389.

Heidemann, D (1996). A queueing theory approach to speed-flow-density relationships. In: *Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, Lyon, France.

Heidemann, D and H Wegmann (1997). Queueing at unsignalized intersections. *Transportation Research Part B*, 31, 239–263.

Heidemann, D (1999). Non-stationary traffic flow from a queueing theory viewpoint. In: *Proceedings of the 14th International Symposium on Transportation and Traffic Theory*, Jerusalem, Israel.

Hongler, M-O and R Filliger (2002). Mesoscopic derivation of a fundamental diagram of one-lane traffic. *Filliger Physics Letters A*, 301, 408–412.

Hopp, WJ and ML Spearman (1996) *Factory Physics, Foundations for Manufacturing Management*. Mc-Graw Hill.

Jain, R and J MacGregor Smith (1997). Modeling vehicular traffic flow using M/G/C/C state dependent queueing models. *Transportation Science*, 31, 324–336.

Jain, S and J Macgregor Smith (1994). Open finite queueing networks with M/M/C/K parallel servers. *Computers Operations Research*, 21(3), 297–317.

Kelly, FP (1979). *Reversibility and Stochastic Networks*. Wiley.

Kerbache, L and J Macgregor Smith (1987). The generalized expansion method for open finite queueing networks. *European Journal of Operational Research*, 32, 448–461.

Kerbache, L and J Macgregor Smith (1988). Assymptotic behavior of the expansion method for open finite queueing networks. *Computurs and Operations Research*, 15(2), 157–169.

Kerbache, L and J MacGregor Smith (2000). Multi-objective routing within large scale facilities using open finite queueing networks. *European Journal of Operational Research*, 121, 105–123.

Kerbache, L and T Van Woensel (2005). Planning and scheduling transportation vehicle fleet in a congested traffic environment. In *Supply Chain Management — European perspectives*. Copenhagen Business School.

Kim, NK and KC Chae (2003). Transfrom-free analysis of the GI/G/1/K queue through the decomposed little's formula. *Computers and Operations Research*, 30(3), 353–365.

Kingman, JFC (1964). The single server queue in heavy traffic. In: *Proceedings of the Cambridge Philosophical Society*, 57, Cambridge University Press, pp. 902–904.

Kleinrock, L (1975). *Queueing Systems: Volume I: Theory*. Wiley & Sons.

Kraemer, W and M Lagenbach-Belz (1976). Approximate formulae for the delay in the queueing system GI/GI/1. In: *Congressbook of the Eight International Teletraffic Congress*, Melbourne, 235–1/8.

Kuehn, PJ (1979). Approximate analysis of general queueing networks by decomposition. *IEEE Trans. Comm*, 27, 113–126.

Leonard, DR, P Power and NB Taylor (1989). CONTRAM: Structure of the model. TRL Report RR 178, Transportation Research Laboratory, Crowthorn.

MacGregor Smith, J and FRB Cruz (2005a). The buffer allocation problem for general finite buffer queueing networks. *IIE Transactions*, 37(4), 343–365.

MacGregor Smith, J and FRB Cruz (2005b). Deterministic and stochastic travel time estimation formulas. Submitted. Available at: ftp://ftp.est.ufmg.br/pub/fcruz/publics/ttime.pdf.

Malek-Madani, R (1997). *Advanced Engineering Mathematics*. Addison Wesley Longman.

Marchal, WG (1976). An approximate formula for waiting time in single server queues. *IIE Transactions*, 8, 473.

Marshall, KT (1968). Some inequalities in queueing. *Operations Research*, 16, 651–665.

May, A (1990). *Traffic Flow Fundamentals*. Prentice-Hall.

May, AD and H Keller (1957). A deterministic queuing model. *Transportation Research*, 1(1), 117–128.

Nelson, P and A Sopasakis (1998). The Prigogine–Herman kinetic model predicts widely scattered traffic flow data at high concentrations. *Transportation Research Part B*, 32(8), 589–604.

Newell, GF (2002). Memoirs on highway traffic flow theory in the 1950s. *Operations Research*, 50(1), 173–178.

Page, E (1972). Queueing theory in OR. *Operational Research Series*.

Perros, HG (1994). *Queueing Networks with Blocking*. Oxford University Press.

Prigogine, I and FC Andrews (1960). A Boltzmann-like approach for traffic flow. *Operations Research*, 8, 789–797.

Qin, L and BL Smith (2001). Characterization of accident capacity reduction. Final report for ITS Center project: Incident capacity estimation.

Richards, PI (1956). Shock waves on the highway. *Operations Research*, 4, 42–51.

Spinellis, D, C Papodopoulos and J Macgregor Smith (2000). Large production line optimisation using simulated annealing. *International Journal of Production Research*, 38(3), 509–541.

Suri, R, JL Sanders and M Kamath (1993). Performance Evaluation of Production Networks. In: *Logistics of Production and Inventory*. North-Holland.

Transportation Research Board (1996). Traffic flow theory: A state-of-the-art report. Technical report, Transporation Research Board.

Transportation Research Board (1998). Highway capacity manual. Technical Report Special Report 209, National Research Council.

Vandaele, N (1996). The impact of lot sizing on queueing delays: Multi product, multi machine models. PhD thesis, Department of Applied Economics, Katholieke Universiteit Leuven.

Vandaele, N, T Van Woensel and A Verbruggen (2000). A queueing based traffic flow model. *Transportation Research Part D*, 5(2), 121–135.

Van Woensel, T (2003). Modeling uninterrupted traffic flows, A queueing approach. Ph.D. Dissertation, University of Antwerp, Belgium.

Van Woensel, T and N Vandaele (2006). Empirical validation of a queueing approach to uninterrupted traffic flows. *4OR, A Quarterly Journal of Operations Research*, 4(1), 59–72.

Van Woensel, T, B Wuyts and N Vandaele (2006). Validating state-dependent queueing models for uninterrupted traffic flows using simulation. *4OR, A Quarterly Journal of Operations Research*, 4(2), 159–174.

Van Woensel, T, B Wuyts and N Vandaele (2005). A queueing theory approach to congestion costs. Submitted.

Van Woensel, T, L Kerbache, H Peremans and N Vandaele (2007). A queueing framework for routing problems with time-dependent travel times. *Journal of Mathematical Modeling and Algorithms*, 6(1), 151–173.

Walrand, J (1988). *An Introduction to Queueing Networks*. Prentice-Hall, Englewood Cliffs.

Whitt, W (1984). Open and closed models for networks of queues. *AT and T Bell Laboratories Technical Journal*, 63(9), 1911–1979.

Whitt, W (1993). Approximations for the GI/G/m queue. *Production and Operations Management*, 2(2), 114–161.

Wuyts, B, T Van Woensel and N Vandaele (2004). State dependent queueing models for traffic. Technical report, Faculty of Applied Economics, University of Antwerp.

Zhang, HM and T Kim (2005). A car-following theory for multiphase vehicular traffic flow. *Transportation Research Part B*, 39, 385–399.

Zijm, WHM (2002). *Manufacturing and Logistic Systems Analysis, Planning and Control*. Universiteit Twente.

Tom van Woensel is an Assistant Professor of Operations Management at the Eindhoven University of Technology (The Netherlands). He teaches at the Department of Industrial Engineering at both undergraduate and graduate levels. He received his PhD degree from the Department of Applied Economic Sciences at the University of Antwerp (Belgium). Tom's main research interests are currently: Retail Operations, Traffic Modeling, Vehicle Routing Problems (building on the queueing models for traffic flows) and Performance Evaluation of Manufacturing Systems (via queueing models). More information with regards to his publications and collaborations can be found on http://home.tm.tue.nl/tvwoense/index.html.

Nico Vandaele holds a degree in Commercial Engineering and a PhD from the KU Leuven. He is a Full Professor of Operations Management at the Universities of Antwerp and Leuven, where he teaches Supply Chain Management, Production Management, Project Management and Performance Analysis of Manufacturing Systems. Research interests are in planning systems, factory physics and traffic modeling. He has published in leading journals like *IIE Transactions, Management Science, Transportation Research, European Journal of Operational Research, and Interfaces*, among others. He is active in several executive training programs and is consultant for many companies, like Inbev, Atlas Copco, IBM, Baxter, Johnson and Johnson.