# Threshold queueing describes the fundamental diagram of uninterrupted traffic

Niek Baer, Richard J. Boucherie, Jan-Kees van Ommeren,

Stochastic Operations Research, Department of Applied Mathematics, University of Twente,

Drienerlolaan 5, 7500 AE Enschede, The Netherlands

**Abstract**

Queueing due to congestion is an important aspect of road traffic. This paper provides a brief overview of queueing models for traffic and a novel threshold queue that captures the main aspects of the empirical shape of the fundamental diagram. Our numerical results characterises the sources of variation that influence the shape of the fundamental diagram.

**Keywords:** Threshold queue, Hysteresis, Capacity drop, Fundamental diagram, Matrix analytic methods, Level dependent quasi-birth-and-death process

## 1 Introduction

Greenshields [13] captures the empirical relation between speed, flow and density for uninterrupted traffic in the fundamental diagram, see Figure 1. Mathematical models for uninterrupted traffic have been developed and the fundamental diagram in its basic form is now well-understood, see e.g. Newell [29, 30, 31] for a concise exposition.

Traffic jams are a major concern for highway operation and may occur in high density traffic due to variability in driving speed. A wide range of traffic models has been developed over the past decades. These models are mainly from statistical physics and non-linear dynamics, see [9, 18]. Congestion due to variable arrival and/or service processes is the main topic of queueing theory, that, however, has hardly been invoked to analyse the fundamental concepts of uninterrupted traffic flows. Notable exceptions are the models introduced by Heidemann [14] and Jain and Smith [21]. However, these models do not capture the

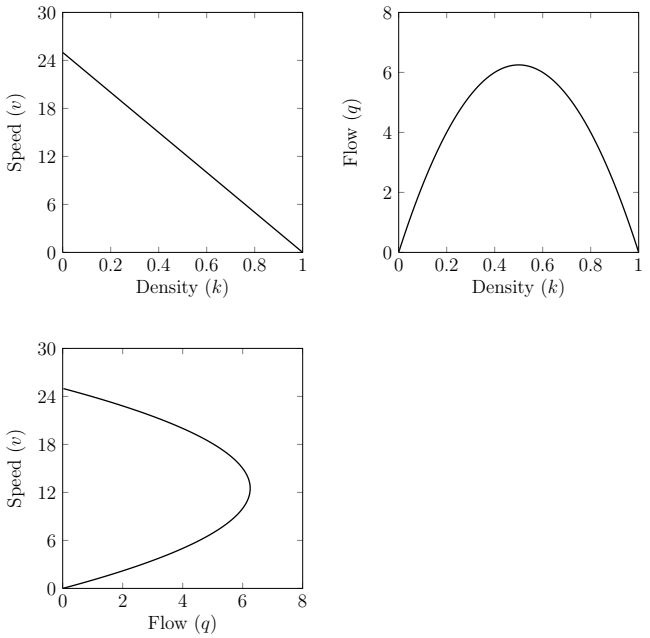


Figure 1: Fundamental diagram from the experimental data of Greenshields [13]

empirical shape of the fundamental diagram for modern traffic as shown in Figure 2. This paper introduces the so-called threshold queue to model and capture the fundamental diagram.

Customers arrive to the threshold queue and require service from a server. Both the interarrival times and service times are controlled by a threshold policy consisting of a lower threshold $L$ and an upper threshold $U$. The empty queue corresponds to a non-congested state in which the server has a high service speed. When the queue length exceeds $U$, the queue reaches a congested state in which the server switches to a low service speed. The queue will switch back to the non-congested state with high service rate once the queue length drops below $L$. Typically $L < U$ which mimics the behaviour on highways where it takes some time for drivers to resume speed.

An important aspect of modern traffic flows is the *capacity drop*, the sharp descent in the fundamental diagram, see Figure 2. In [18], Helbing explains the capacity drop as the transition from non-congested traffic to congested traffic. When the density of vehicles reaches a certain critical value, $\rho_2$, traffic will become congested and the average speed is significantly lower than in non-congested traffic. When density decreases and reaches another critical value, $\rho_1 \leq \rho_2$, a transition from congested traffic to non-congested traffic occurs and traffic flows recover. In the density interval $[\rho_1, \rho_2]$ both congested and non-congested traffic flows exist, which indicates the existence of hysteresis. As is shown in our numerical results, it is

precisely this hysteresis effect captured by the threshold queue that results in the capacity drop in the fundamental diagram of Figure 2 observed in empirical data for speed, flow and density.

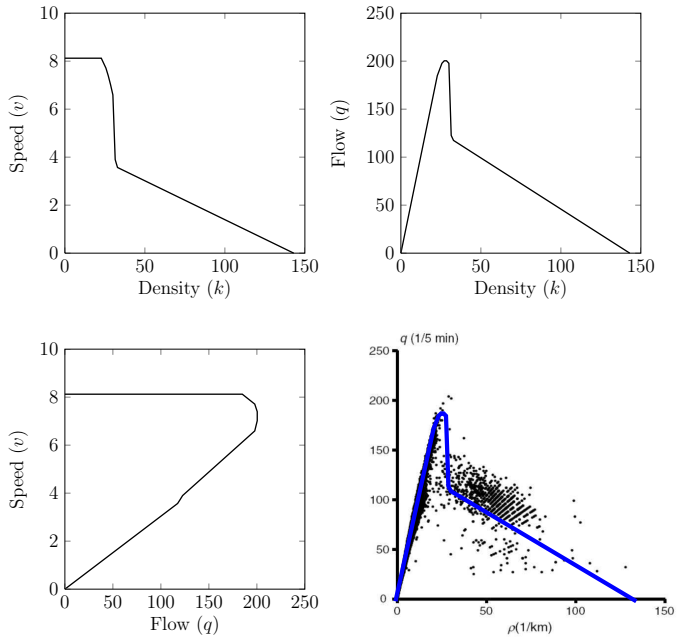


Figure 2: Fundamental diagram from experimental data for modern traffic [37]. The flow-density diagram is fitted to the experimental data.

Section 2 gives a brief overview of the literature on queueing models for uninterrupted traffic flows. The threshold queue model is introduced and analysed in Section 3. Results on the fundamental diagram obtained with the threshold queue are shown in Section 4. Section 5 gives concluding remarks.

## 2 Literature

Congestion is a key concept in queueing theory that models both the mesoscopic and macroscopic effects of randomness on delay and sojourn times. In interrupted traffic flows, where queues arise naturally at an intersection, queueing theory has been a popular tool since the early 1940s, see [5] for a recent survey. In uninterrupted traffic, however, queueing models have received far less attention in literature. In this section we focus on queueing models for uninterrupted traffic flows.

## 2.1 Microscopic, mesoscopic and macroscopic models

Uninterrupted or highway traffic flow models can be characterised by their level of detail: microscopic, mesoscopic and macroscopic.

In *microscopic models* a high level of detail is used in which each individual driver is characterised by its position and behaviour over time [19, 32]. In general, microscopic models lead to systems of (ordinary) differential equations [18]. Well-known micropscopic models are the car-following model [6, 12], the cellular automata model [27] and lane-changing models [1].

In *mesoscopic models* the individual drivers are not distinguished [19, 32]. The behaviour of drivers is characterised in terms of the probability density $f(x, v, t)$ of vehicles at position $x$ with speed $v$ at time $t$. Examples of mesoscopic models are headway distribution models [8] and gas-kinetic continuum models [34, 35].

*Macroscopic models* have the lowest level of detail and consider only three variables for each position $x$ and time $t$: average speed $v(x, t)$, traffic flow $q(x, t)$ and spatial vehicle density $k(x, t)$, that are related as $q(x, t) = v(x, t) \cdot k(x, t)$. These three variables are often presented in the fundamental diagram. Two classical examples of macroscopic models are the Lighthill-Whitham-Richards models [25, 26, 36] and the Payne models [33].

For more elaborate surveys on traffic models for uninterrupted traffic flows, see [4, 18, 19, 32, 40].

## 2.2 Queueing Theory in uninterrupted traffic flows

Two main queueing theoretic approaches can be identified to model uninterrupted traffic: the queue with waiting room of Heidemann [14] and the queue with blocking of Jain and Smith [21]. We use the following notation. Let $k$ denote the traffic density, $v$ the mean speed of a vehicle, $q$ the flow rate, $k_{jam}$ the jam or maximum density and $v_f$ the desired mean speed or free flow speed.

**Heidemann's model.** Heidemann [14] introduces an $M/G/1$ queueing system to model highway traffic. The server in the queueing system corresponds to a highway segment of length $1/k_{jam}$, which is the minimal part of the highway each vehicle requires. The mean service time in the queue is the average time it takes a vehicle in free flow traffic to cross the segment: $\mathbb{E}[B] = 1/(k_{jam} \cdot v_f)$. The traffic density outside the chosen segment is $k$, so that the mean time between two arrivals is $\mathbb{E}[A] = 1/(k \cdot v_f)$. In the $M/G/1$ queue, an arriving vehicle may find the server busy upon arrival and must wait for service. The total time required to cross the segment is the sojourn time, $\mathbb{E}[S]$, which is the sum of the waiting time and the service time. For

the $M/G/1$ queue the Pollaczek-Khintchine formula [43] gives:

$$\mathbb{E}[S] = \mathbb{E}[B]\left[1 + \frac{\rho}{1-\rho} \cdot \frac{(1+c_s^2)}{2}\right],$$

where $c_s$ is the coefficient of variation of the service time. The speed, $v$, of a vehicle passing the segment then is

$$v = \frac{1/k_{jam}}{\mathbb{E}[S]}.$$

Figure 3 gives the fundamental diagram for the $M/G/1$ queue for various choices of $c_s$.

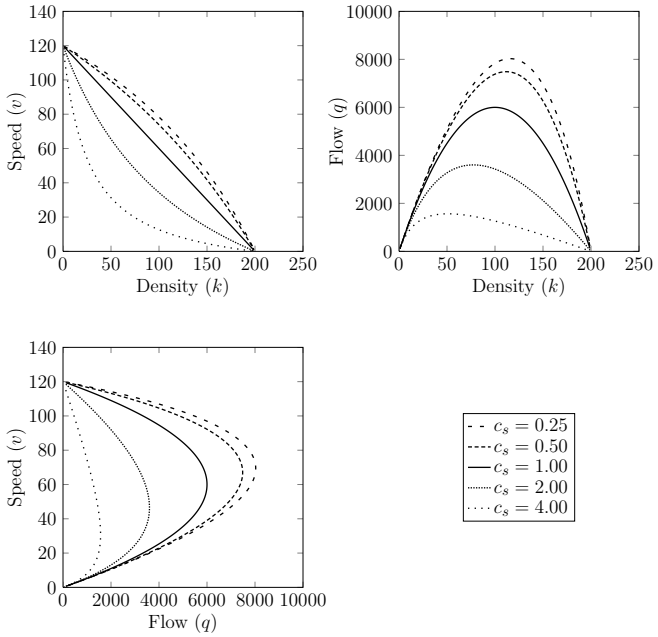


Figure 3: Fundamental diagram obtained with Heidemann's $M/G/1$ queue, $k_{jam} = 200, v_f = 120$ and varying coefficient of variation $c_s$.

Generalisations of Heidemann's model include the transient analysis of the $M/G/1$ queue in [15, 16, 17]. Vandaele, Van Woensel and Verbruggen [42] and Van Woensel [38] consider the $G/G/s$ queue. Validation of their queueing model [39, 41] shows that the $M/G/1$ queue models non-congested traffic and the $G/G/s$ is a more suitable model for congested traffic. Accidents were incorporated by Baykal-Gürsoy, Xiao and Ozbay [3] in an $M/MSP/c$ queue with service rates represented by a Markovian Service Process.

**Jain and Smith's Model.** An alternative approach to a queueing model for highway traffic is the $M/G/c/c$ model of Jain and Smith [21], where an arrival that finds all servers occupied is blocked and

cleared (lost). Their model is based on pedestrian flows in emergency evacuation planning as introduced in [44]. The servers correspond to a road segment. As the $M/G/c/c$ model does not incorporate waiting, the speed of a vehicle is obtained by the service time that equals the sojourn time in the queue. The capacity $C$ of a road segment equals the number of vehicles that fit in this segment, i.e., the product of the jam density, $k_{jam}$, the length of the road segment, $L$, and the number of lanes, $N$: $C = k_{jam} \cdot L \cdot N$. The mean speed of a vehicle, $V_n$, depends on the number of vehicles $n$ on the road segment and is now a function that is input for the model. Two functions for $V_n$ are considered in [21, 44]:

$$V_n = \frac{v_f}{C}\left(C + 1 - n\right),$$

that linearly decreases in the number of vehicles on the segment, and

$$V_n = v_f \cdot \exp\left[-\left(\frac{n-1}{\beta}\right)^{\gamma}\right],$$

for suitable constants $\gamma$ and $\beta$, see [44], that exponentially decreases with the number of vehicles. In Figure 4 we present the fundamental diagram obtained with the $M/G/c/c$ queue for both speed functions $V_n$.

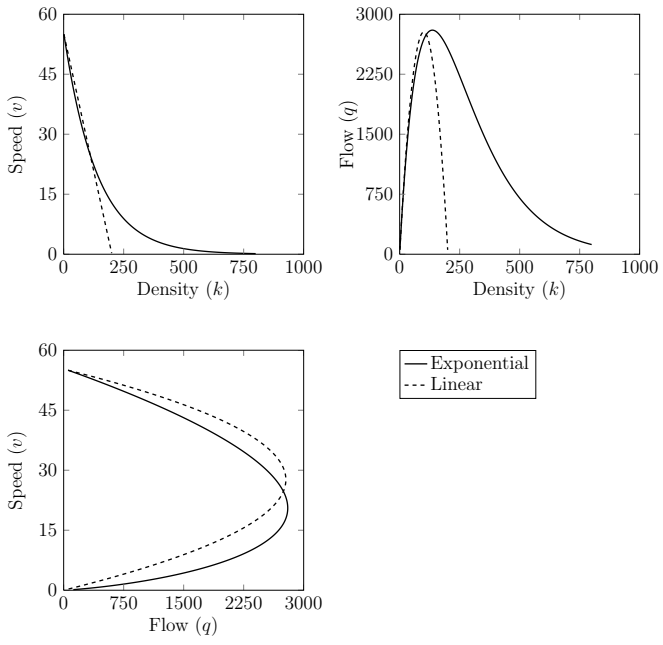


Figure 4: Fundamental diagram obtained with Jain and Smith's $M/G/c/c$ queue for a linear and exponential decreasing speed, $k_{jam} = 200, v_f = 55$ mph.

A network of $M/G/c/c$ queues was considered in Cruz, Smith and Medeiros [11] and Cruz and Smith [10]. In this network a blocked customer will occupy its server until it is no longer blocked. In [10, 11] simulation techniques and approximations were used to derive blocking probabilities, throughput, mean queue length and mean waiting times.

**Summary.** The queueing models in literature result in a fundamental diagram similar to the fundamental diagram by Greenshields. However, they do not capture the hysteresis effect as seen in modern traffic flows. The threshold queue in the next section mimics this hysteresis effect and will be shown to capture the resulting capacity drop.

# 3 Threshold queue with hysteresis

Consider a single server queue where service rates are controlled by a threshold policy. Customers arrive according to a Poisson process with rate $\lambda$ and require an exponential service time, depending on the stage of the queue. This stage is either non-congested (denoted by stage 1) or congested (denoted by stage 2) and is controlled by the threshold policy. Once an arrival occurs while the queue length is $U$, the stage changes from non-congested to congested. The stage changes back form congested to non-congested when the queue length is $L$ and a departure occurs. The state space of this threshold queue is depicted in Figure 5. The stationary queue length probabilities $\pi$ for this threshold queue can readily be obtained from standard

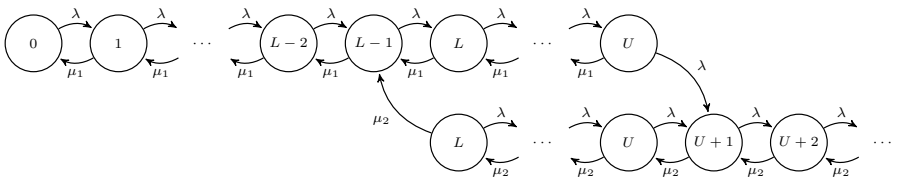


Figure 5: State Diagram for the $M/M/1$ threshold queue.

Markov Chain analysis, see also Le Ny and Tuffin [24]. Let $\rho = \frac{\lambda}{\mu_1}$ and $\delta = \frac{\lambda}{\mu_2}$, then $\pi_{i,1}$ and $\pi_{i,2}$, where

$\pi_{i,j}$ denotes the probability of having $i$ customers in the queue in stage $j$, are given by:

$$\begin{aligned}
\pi_{i,1} &= \rho^i \pi_0, && i = 1, \ldots, L-1, \\
\pi_{i,1} &= \frac{\rho^i - \rho^{U+1}}{1-\rho^{U-L+2}} \pi_0, && i = L, \ldots, U, \\
\pi_{i,2} &= \frac{\delta - \delta^{i-L+2}}{1-\delta} \frac{\rho^U - \rho^{U+1}}{1-\rho^{U-L+2}} \pi_0, && i = L, \ldots, U, \\
\pi_{i,2} &= \frac{\delta^{i-U+2} - \delta^{i-L+2}}{1-\delta} \frac{\rho^U - \rho^{U+1}}{1-\rho^{U-L+2}} \pi_0, && i = U, U+1, \ldots,
\end{aligned}$$

with

$$\pi_0 = \left[ \frac{(1-\rho^{U-L+2})(1-\delta) - \rho^U (U-L+2)(\rho-\delta)(1-\rho)}{(1-\rho^{U-L+2})(1-\rho)(1-\delta)} \right]^{-1}.$$

The mean sojourn time is:

$$\begin{aligned}
\mathbb{E}[S] = \frac{\pi_0}{\lambda} \Bigg[ & \frac{\rho}{(1-\rho)^2} + \left( \frac{(U-L+2)\rho^U}{1-\rho^{U-L+2}} \right) \left( \frac{\delta(1-\rho)(\delta - 2\rho\delta + 2\rho^2) + \rho^2(\rho + U + L)}{(1-\delta)^2(1-\rho)^2} \right) \\
& + \left( \frac{((U+1)(U+2) - L(L-1))\rho^U}{1-\rho^{U-L+2}} \right) \\
& \cdot \left( \frac{\delta(1-\rho)(1-\delta) - \rho(1-\delta)(1+\delta+\rho^2) - \rho^2\delta(1+\delta)}{(1-\delta)^2(1-\rho)^2} \right) \Bigg],
\end{aligned}$$

from which the fundamental diagram can be obtained using Heidemann's model.

For traffic modelling, the assumption of Poisson arrivals and exponential service are far from realistic. Therefore, we consider the $PH/PH/1$ threshold queue below in which both the interarrival times and service times are of phase-type. The phase-type distribution allows us to approximate any distribution with non-negative support arbitrarily close, see [20]. A phase-type distribution is described by an absorbing Markov Chain consisting of $n$ transient states and 1 absorbing state and generator

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{A}^0 \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix},$$

where $\boldsymbol{A}$ is an $n \times n$ matrix describing transition rates between transient states and $\boldsymbol{A}^0$ an $n \times 1$ vector describing the transition rates into the absorbing state. Let $\boldsymbol{e}_n$ denote the $n \times 1$ vector of all ones, then $\boldsymbol{A}^0 = -\boldsymbol{A}\boldsymbol{e}_n$. The initial state probability vector of $\boldsymbol{G}$ is $\boldsymbol{g} = [\boldsymbol{\alpha}, \alpha_{n+1}]$ with $\boldsymbol{\alpha}\boldsymbol{e} + \alpha_{n+1} = 1$. Here we

assume that $\alpha_{n+1} = 0$ such that the absorbing state is never chosen as initial state. We denote a phase-type distribution with generator $\boldsymbol{G}$ and probability vector $\boldsymbol{g}$ by $PH(\boldsymbol{A}, \boldsymbol{\alpha})$.

Consider the $PH/PH/1$ threshold queue with interarrival times and service times having phase-type ($PH$) distribution. In the $PH/PH/1$ threshold queue the interarrival times are $PH(\boldsymbol{\Lambda}, \boldsymbol{\lambda})$ distributed and the service times are either $PH(\boldsymbol{M}_1, \boldsymbol{\mu}_1)$ (non-congested) or $PH(\boldsymbol{M}_2, \boldsymbol{\mu}_2)$ (congested) distributed. We assume that the queue is stable, the mean service time in the congested stage is less than the mean interarrival time. Furthermore, we assume that the mean service time in the congested stage is larger than the mean service time in the non-congested stage.

The $PH/PH/1$ threshold queue is a four-dimensional Markov Chain $(n, s, x, y)$, where $n$ represents the queue length, $s$ the stage of the queue, $x$ the state of the arrival process and $y$ the state of the service process. We model this queue as a Level Dependent Quasi-Birth-and-Death (LDQBD) process in which the levels are formed by the queue length [23]. The phases are formed by the stage of the system and the states of the service and arrival distributions. We use Matrix Analytic Methods, see [7, 23], to obtain the stationary queue length distribution. The generator $\boldsymbol{Q}$ of a LDQBD has the following tri-diagonal structure:

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{L}^{(0)} & \boldsymbol{F}^{(0)} & 0 & \cdots & & \\ \boldsymbol{B}^{(1)} & \boldsymbol{L}^{(1)} & \boldsymbol{F}^{(1)} & \ddots & & \\ 0 & \boldsymbol{B}^{(2)} & \boldsymbol{L}^{(2)} & \ddots & & \\ \vdots & \ddots & \ddots & \ddots & \boldsymbol{F}^{(i-1)} & \\ & & & \boldsymbol{B}^{(i)} & \boldsymbol{L}^{(i)} & \ddots \\ & & & & \ddots & \ddots \end{bmatrix}.$$

Here $\boldsymbol{F}^{(i)}$ describes the *forward* transitions from level $i$ to level $i+1$ (arrivals), $\boldsymbol{L}^{(i)}$ the *local* transition within level $i$ and $\boldsymbol{B}^{(i)}$ the *backward* transitions from level $i$ to level $i-1$ (departures).

Let $\boldsymbol{\pi} = [\boldsymbol{\pi_0}, \boldsymbol{\pi_1}, \ldots]$ denote the probability vector such that $\boldsymbol{\pi Q} = \boldsymbol{0}$ and $\boldsymbol{\pi e} = 1$. The elements of the probability vector $\boldsymbol{\pi_i}$ describe the probability of being in a certain phase and in level $i$, i.e. the probability of having $i$ customers in the queue. If the LDQBD is irreducible, aperiodic and positive recurrent, the stationary queue length distribution $\boldsymbol{\pi}$ is given by:

$$\boldsymbol{\pi_i} = \boldsymbol{\pi_0} \prod_{n=0}^{i-1} \boldsymbol{R}^{(n)},$$

where $\boldsymbol{\pi_0}$ is subject to the boundary conditions:

$$\boldsymbol{\pi_0}\left(\boldsymbol{L}^{(0)} + \boldsymbol{R}^{(0)}\boldsymbol{B}^{(1)}\right) = \mathbf{0},$$

and to the normalisation conditions:

$$\boldsymbol{\pi_0}\left(\sum_{i=0}^{\infty}\prod_{n=0}^{i-1}\boldsymbol{R}^{(n)}\right)\boldsymbol{e} = 1.$$

The matrices $\boldsymbol{R}^{(n)}$ are the minimal non-negative solution to the set of equations

$$\boldsymbol{F}^{(i)} + \boldsymbol{R}^{(i)}\boldsymbol{L}^{(i+1)} + \boldsymbol{R}^{(i)}\boldsymbol{R}^{(i+1)}\boldsymbol{B}^{(i+2)} = \mathbf{0}, \quad i \leq 0.$$

Let $\mathbb{E}[A]$ denote the mean interarrival time and $\mathbb{E}[S]$ the mean sojourn time. We compute the mean queue length using $\boldsymbol{\pi}$ and determine the mean sojourn time using Little's Law:

$$\mathbb{E}[S] = \mathbb{E}[A]\left[\sum_{n=0}^{\infty} n\boldsymbol{\pi_n}\boldsymbol{e}\right]$$

Appendix A gives a detailed description of the generator $\boldsymbol{Q}$ and the procedure described in this section.

# 4 Sensitivity Analysis

In this section we numerically investigate the fundamental diagram for the threshold queue. Figure 6 shows the fundamental diagram for the $M/M/1$ threshold queue with $k_{jam} = 1, \mu_1 = 25, \mu_2 = 15, L = 5$ and $U = 10$. The capacity drop appears at a density of 0.6.

Below we perform a sensitivity analysis on the influence of the parameters $L, U, \mu_1$ and $\mu_2$ in the $M/M/1$ threshold queue on the shape of the fundamental diagram and on the distribution functions in the $PH/PH/1$ threshold queue (while leaving $L, U, \mu_1$ and $\mu_2$ unchanged).

**$M/M/1$ threshold queue** Figure 7 characterises the fundamental diagram of the $M/M/1$ threshold queue for four different scenarios. Each scenario is based on the scenario of Figure 6 but in each subgraph one of the four parameters is altered: (a) the lower threshold, $L$, (b) the upper threshold, $U$, (c) the high service rate, $\mu_1$ and (d) the low service rate, $\mu_2$.

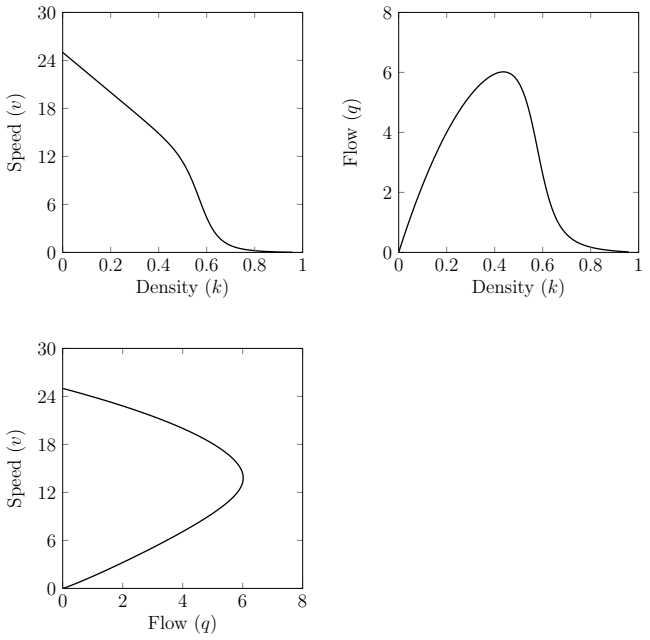


Figure 6: Fundamental diagram for the $M/M/1$ threshold queue.

The effects of altering $L$ are minimal, as can be seen in Figure 7(a). Figure 7(b) shows that the steepness of the capacity drop increases by increasing $U$. Figures 7(c) and 7(d) show that the position of the capacity drop varies with $\mu_1$ or $\mu_2$ and that flows increase with $\mu_1$.

$PH/PH/1$ **threshold queue** Figure 8 characterises the fundamental diagram of the $PH/PH/1$ threshold queue for four different scenarios. We select phase-type distributions such that the mean inter-arrival times and mean service times (in both stages) are the same as in the $M/M/1$ threshold queue of Figure 6. We study three different distributions, the Hyper-Exponential distribution with four phases, $H_4$, the Exponential distribution, $M$, and the Erlang distribution with 4 phases, $E_4$, with respectively $c^2_{H_4} = 1.5744, c^2_M = 1$, and $c^2_{E_4} = 0.25$. Furthermore, we set $L = 5$ and $U = 15$.

In Figure 8(a) we vary $PH(\boldsymbol{M}_2, \boldsymbol{\mu}_2)$, the distribution of the congested service process. We set $PH(\boldsymbol{\Lambda}, \boldsymbol{\lambda}) = H_4$ and $PH(\boldsymbol{M}_1, \boldsymbol{\mu}_1) = H_4$. It can be seen that the fundamental diagrams for the three different distributions of $PH(\boldsymbol{M}_2, \boldsymbol{\mu}_2)$ are similar. This implies that the coefficient of variation of $PH(\boldsymbol{M}_2, \boldsymbol{\mu}_2)$ has minor influence on the fundamental diagram.

In Figure 8(b) we vary $PH(\boldsymbol{\Lambda}, \boldsymbol{\lambda})$, the distribution of the arrival process. We set $PH(\boldsymbol{M}_1, \boldsymbol{\mu}_1) = H_4$ and $PH(\boldsymbol{M}_2, \boldsymbol{\mu}_2) = H_4$. We observe that increasing variability in the arrival process reduces speed and flow for

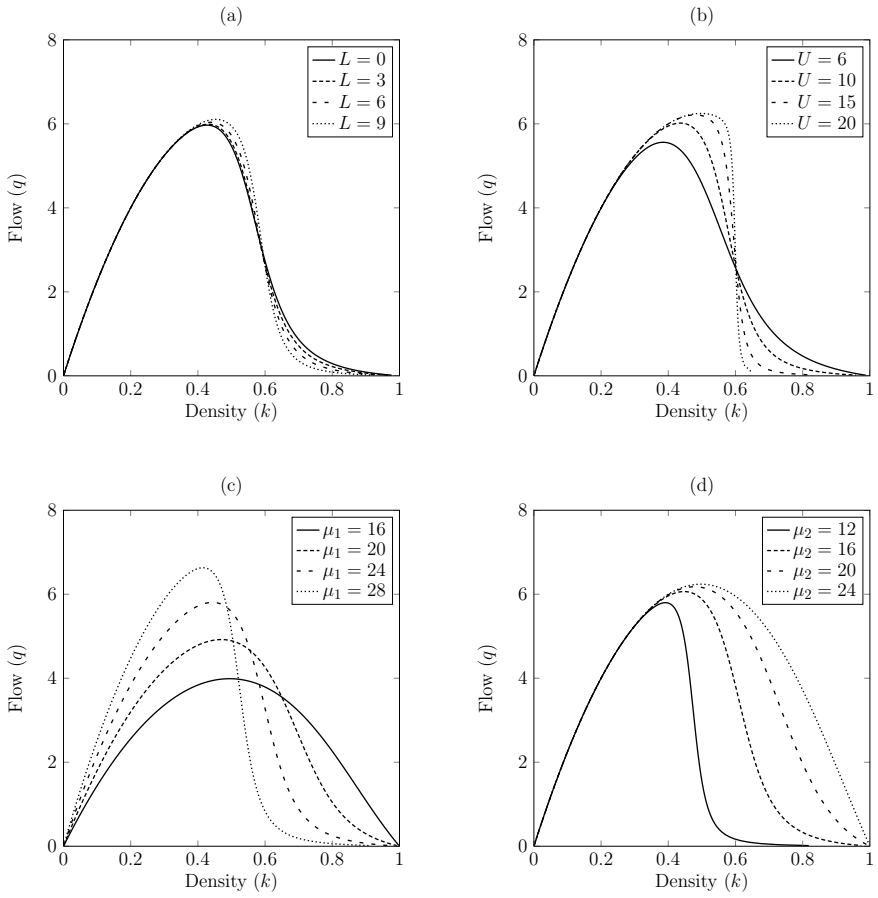


Figure 7: Flow-Density diagram for the $M/M/1$ threshold queue with varying (a) $L$, (b) $U$, (c) $\mu_1$ and (d) $\mu_2$.

density $k < 0.6$. For $k > 0.6$, flows and speeds increase when the variability decreases.

In Figure 8(c) we vary $PH(\boldsymbol{M}_1, \boldsymbol{\mu}_1)$, the distribution of the non-congested service process. We set $PH(\boldsymbol{M}_1, \boldsymbol{\mu}_1) = H_4$ and $PH(\boldsymbol{\Lambda}, \boldsymbol{\lambda}) = E_4$. Here the effects of the Erlang distribution are clearly visible in the fundamental diagram. In the case where both the arrival process and the non-congested service process are Erlang with four phases, the fundamental diagram reaches a maximum density of 0.6. This is a result of the low probability of reaching the congested stage caused by the low variability in the $E_4$ distribution. Note that for deterministic arrival and service processes the congested stage is never reached. The maximum density of 0.6 is obtained for a mean interarrival time close to $1/15$, the mean service time in the congested stage. For this value the queue is still stable and the density is $15/25 = 0.6$.

In Figure 8(d) we vary $PH(\boldsymbol{M}_1, \boldsymbol{\mu}_1)$, the distribution of the non-congested service process. We set $PH(\boldsymbol{M}_1, \boldsymbol{\mu}_1) = H_4$ and $PH(\boldsymbol{\Lambda}, \boldsymbol{\lambda}) = H_4$. The results are similar to those for Figure 8(b). Increasing

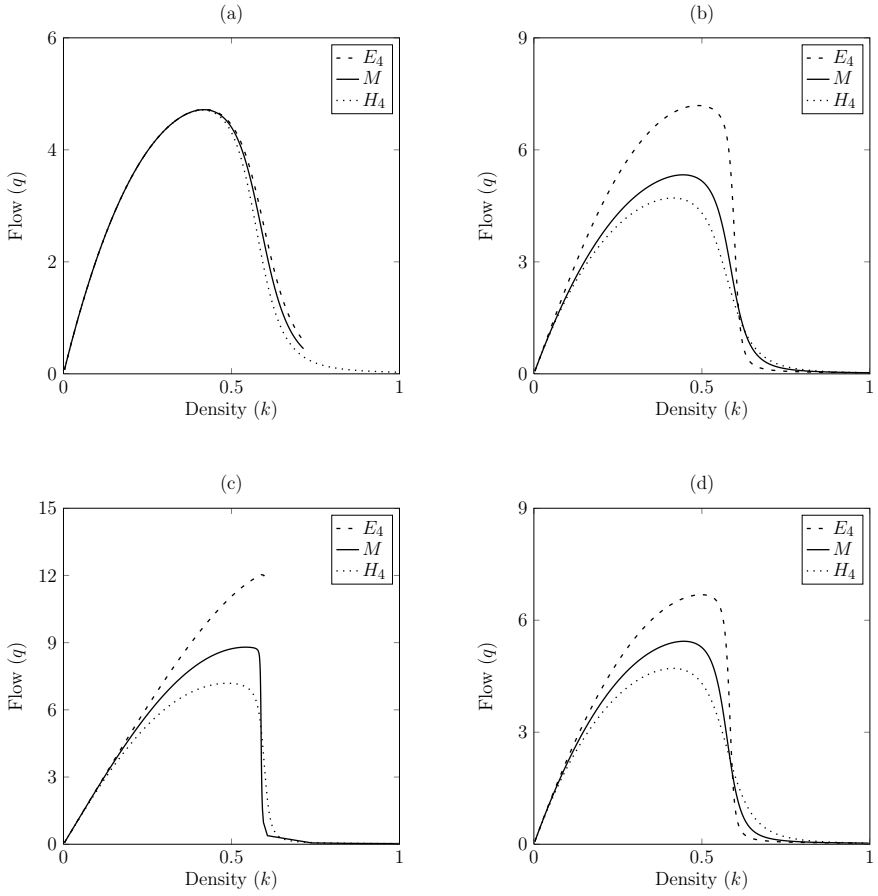


Figure 8: Flow-Density diagram for the $PH/PH/1$ threshold queue by selecting various distributions for the (a) congested service process, (b) arrival process, (c) and (d) non-congested service process.

variability of the non-congested service process decreases flow and speeds for $k < 0.6$. For $k > 0.6$, flows and speeds increase when the variability decreases.

# 5 Conclusions

This paper has introduced the $PH/PH/1$ threshold queue to study the parameters of traffic that influence the shape of the fundamental diagram including the capacity drop in this diagram observed in empirical data for modern traffic flows.

The $PH/PH/1$ threshold queue has two service regimes: high and low service rates, and switches from high rates to low rates when the queue length exceeds the upper threshold, and returns to high rates when the queue length falls below the lower threshold, where the lower threshold is smaller than the upper threshold.

Sensitivity analysis reveals that steepness of the capacity drop and traffic density where it occurs are determined by respectively the value of the higher threshold and the mean service times.

The service distribution in the congested stage has minor influence on the fundamental diagram. In contrast, increasing variability in the arrival process or non-congested service process is shown to reduce both speed and flow for a density less than the capacity drop density.

# A $PH/PH/1$ Threshold queue

A phase-type distribution is described by an absorbing Markov Chain consisting of $n$ transient states and 1 absorbing state and generator

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{A}^0 \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix},$$

where $\boldsymbol{A}$ is a $n \times n$ matrix describing transition rates between transient states and $\boldsymbol{A}^0 = -\boldsymbol{A}\boldsymbol{e}$ an $n \times 1$ vector describing the transition rates into the absorbing state. The initial state probability vector of $\boldsymbol{G}$ is $\boldsymbol{g} = [\boldsymbol{\alpha}, \alpha_{n+1}]$ with $\boldsymbol{\alpha}\boldsymbol{e} + \alpha_{n+1} = 1$. Here we assume that $\alpha_{n+1} = 0$ such that the absorbing state is never chosen as initial state. We denote a phase-type distribution with generator $\boldsymbol{G}$ and probability vector $\boldsymbol{g}$ by $PH(\boldsymbol{A}, \boldsymbol{\alpha})$. The mean time until absorption for the $PH(\boldsymbol{A}, \boldsymbol{\alpha})$ is given by $-\boldsymbol{\alpha}\boldsymbol{A}^{-1}\boldsymbol{e}_n$, where $\boldsymbol{e}_n$ denotes an $n \times 1$ vector of ones.

Consider the $PH/PH/1$ threshold queue with interarrival times and service times having phase-type ($PH$) distribution. In the $PH/PH/1$ threshold queue the interarrival times are $PH(\boldsymbol{\Lambda}, \boldsymbol{\lambda})$ distributed, with $p$ states, and the service times are either $PH(\boldsymbol{M}_1, \boldsymbol{\mu}_1)$ (non-congested), with $q_1$ states, or $PH(\boldsymbol{M}_2, \boldsymbol{\mu}_2)$ (congested), with $q_2$ states, distributed. We assume that the queue is stable and that the mean service time in the non-congested stage is less than the mean service times in the congested stage:

$$-\boldsymbol{\mu}_1\boldsymbol{M}_1^{-1}\boldsymbol{e}_{q_1} < -\boldsymbol{\mu}_2\boldsymbol{M}_2^{-1}\boldsymbol{e}_{q_2} < -\boldsymbol{\lambda}\boldsymbol{\Lambda}^{-1}\boldsymbol{e}_p.$$

The resulting queueing system is a four-dimensional Markov Chain $(n, s, x, y)$, where $n$ represents the queue length, $s$ the stage of the queueing system, $x$ the state of the arrival process and $y$ the state of the service process. This threshold queue is modelled as a Level Dependent Quasi-Birth-and-Death (LDQBD) process in which the levels are formed by the queue length [23]. The phases are formed by the stage of the system and the states of the service and arrival distributions. We order the states of the Markov Chain

lexicographically:

- Level 0: $(0,1,1,1),\ldots,(0,1,1,q_1),\ldots,(0,1,p,1),\ldots,(0,1,p,q_1),\ldots,$ $(0,2,1,1),\ldots,(0,2,1,q_2),\ldots,(0,2,p,1),\ldots,(0,2,p,q_2)$.
- Level 1: $(1,1,1,1),\ldots,(1,1,1,q_1),\ldots,(1,1,p,1),\ldots,(1,1,p,q_1),\ldots,$ $(1,2,1,1),\ldots,(1,2,1,q_2),\ldots,(1,2,p,1),\ldots,(1,2,p,q_2)$.

The generator $\boldsymbol{Q}$ of a LDQBD had the following tri-diagonal structure:

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{L}^{(0)} & \boldsymbol{F}^{(0)} & 0 & \cdots & & \\ \boldsymbol{B}^{(1)} & \boldsymbol{L}^{(1)} & \boldsymbol{F}^{(1)} & \ddots & & \\ 0 & \boldsymbol{B}^{(2)} & \boldsymbol{L}^{(2)} & \ddots & & \\ \vdots & \ddots & \ddots & \ddots & \boldsymbol{F}^{(i-1)} & \\ & & & \boldsymbol{B}^{(i)} & \boldsymbol{L}^{(i)} & \ddots \\ & & & & \ddots & \ddots \end{bmatrix}.$$

Here $\boldsymbol{F}^{(i)}$ describes the *forward* transitions from level $i$ to level $i+1$ (arrivals), $\boldsymbol{L}^{(i)}$ the *local* transitions within level $i$ and $\boldsymbol{B}^{(i)}$ the *backward* transitions from level $i$ to level $i-1$ (departures). Each $\boldsymbol{F}^{(i)}, \boldsymbol{L}^{(i)}$ and $\boldsymbol{B}^{(i)}$ can be denoted by an $2 \times 2$-matrix of submatrices $\boldsymbol{F}^{(i)}_{(j,k)}, \boldsymbol{L}^{(i)}_{(j,k)}$ and $\boldsymbol{B}^{(i)}_{(j,k)}$ respectively, describing the forward, local and backward transition rates from stage $j$ to stage $k$.

Let $\boldsymbol{I}_t$ denote the $t \times t$ identity matrix and let $\otimes$ denote the Kronecker product. The forward, local and

backwards transitions are given by:

$$\begin{aligned}
\boldsymbol{F}^{(i)}_{(1,1)} &= \boldsymbol{\Lambda}^0 \otimes \boldsymbol{\lambda} \otimes \boldsymbol{I}_{q_1}, && 0 \leq i < U, \\
\boldsymbol{F}^{(i)}_{(2,2)} &= \boldsymbol{\Lambda}^0 \otimes \boldsymbol{\lambda} \otimes \boldsymbol{I}_{q_2}, && L \leq i, \\
\boldsymbol{F}^{(U)}_{(1,2)} &= \boldsymbol{\Lambda}^0 \otimes \boldsymbol{e}_{q_1} \otimes \boldsymbol{\lambda} \otimes \boldsymbol{\mu}_2, &&
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{L}^{(i)}_{(1,1)} &= \boldsymbol{\Lambda} \otimes \boldsymbol{I}_{q_1} + \boldsymbol{I}_p \otimes \boldsymbol{M}_1, && 0 < i \leq U, \\
\boldsymbol{L}^{(i)}_{(2,2)} &= \boldsymbol{\Lambda} \otimes \boldsymbol{I}_{q_2} + \boldsymbol{I}_p \otimes \boldsymbol{M}_2, && L \leq i, \\
\boldsymbol{L}^{(0)}_{(1,1)} &= \boldsymbol{\Lambda} \otimes \boldsymbol{I}_{q_1}, &&
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{B}^{(i)}_{(1,1)} &= \boldsymbol{I}_p \otimes \boldsymbol{M}_1^0 \otimes \mu_1, && 0 < i \leq U, \\
\boldsymbol{B}^{(i)}_{(2,2)} &= \boldsymbol{I}_p \otimes \boldsymbol{M}_2^0 \otimes \mu_2, && L < i, \\
\boldsymbol{B}^{(L)}_{(2,1)} &= \boldsymbol{e}_p \otimes \boldsymbol{M}_2^0 \otimes \lambda \otimes \boldsymbol{\mu}_1. &&
\end{aligned}$$

Submatrices of $\boldsymbol{F}$, $\boldsymbol{L}$ and $\boldsymbol{B}$ not defined above are equal to the zero-matrix $\boldsymbol{0}$.

Let $\boldsymbol{\pi} = [\boldsymbol{\pi_0}, \boldsymbol{\pi_1}, \ldots]$ denote the probability vector such that $\boldsymbol{\pi Q} = \boldsymbol{0}$ and $\boldsymbol{\pi e} = 1$. The elements of the probability vector $\boldsymbol{\pi_i}$ describe the probability of being in a certain phase in level $i$. If the LDQBD is irreducible, aperiodic and positive recurrent, the stationary queue length distribution $\boldsymbol{\pi}$ is obtained using the decomposition method in Baer [2] and is given by:

$$\begin{aligned}
\boldsymbol{\pi_i} &= \boldsymbol{\pi_0} \prod_{n=0}^{i-1} \boldsymbol{R}^{(n)}, && 0 < i \leq U+1, \\
\boldsymbol{\pi_i} &= \boldsymbol{\pi_0} \left[ \prod_{j=0}^{U} \boldsymbol{R}^{(j)} \right] \left[ \boldsymbol{R}^{U+1} \right]^{i-U}, && i > U+1,
\end{aligned}$$

in which $\boldsymbol{\pi_0}$ is subject to the boundary conditions:

$$\boldsymbol{\pi_0} \left( \boldsymbol{L}^{(0)} + \boldsymbol{R}^{(0)} \boldsymbol{B}^{(1)} \right) = \boldsymbol{0},$$

and to the normalisation conditions:

$$\boldsymbol{\pi_0}\left(\sum_{i=0}^{\infty}\prod_{n=0}^{i-1}\boldsymbol{R}^{(n)}\right)\boldsymbol{e}=1.$$

The matrices $\boldsymbol{R}^{(i)}$ and $\boldsymbol{R}^{(U+1)}$ are the minimal non-negative solutions to the equations:

$$\boldsymbol{F}^{(i)}+\boldsymbol{R}^{(i)}\boldsymbol{L}^{(i+1)}+\boldsymbol{R}^{(i)}\boldsymbol{R}^{(i+1)}\boldsymbol{B}^{(i+2)}=\boldsymbol{0}, \tag{1}$$

$$\boldsymbol{F}^{(U+1)}+\boldsymbol{R}^{(U+1)}\boldsymbol{L}^{(U+2)}+\boldsymbol{R}^{(U+1)}\boldsymbol{R}^{(U+1)}\boldsymbol{B}^{(U+3)}=\boldsymbol{0}. \tag{2}$$

Decomposing $\boldsymbol{R}^{(i)}$ results in

$$\boldsymbol{R}^{(i)}=\begin{bmatrix}\boldsymbol{R}^{(i)}_{(1,1)} & \boldsymbol{R}^{(i)}_{(1,2)}\\ \boldsymbol{R}^{(i)}_{(2,1)} & \boldsymbol{R}^{(i)}_{(2,2)}\end{bmatrix},$$

and in particular

$$\boldsymbol{R}^{(U+1)}=\begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}^{(U+1)}_{(2,2)}\end{bmatrix}.$$

It follows from (2) that $\boldsymbol{R}^{(U+1)}_{(2,2)}$ is the minimal non-negative solution to

$$\boldsymbol{F}^{(U+1)}_{(2,2)}+\boldsymbol{R}^{(U+1)}_{(2,2)}\boldsymbol{L}^{(U+2)}_{(2,2)}+\boldsymbol{R}^{(U+1)}_{(2,2)}\boldsymbol{R}^{(U+1)}_{(2,2)}\boldsymbol{B}^{(U+3)}_{(2,2)}=\boldsymbol{0}.$$

This matrix equation can be solved numerically by a fixed-point iteration, see [28], or a logarithmic reduction algorithm, see [22]. Knowing $\boldsymbol{R}^{(U+1)}$ one can use (1) to iteratively find all $\boldsymbol{R}^{(i)}$'s. For the

$PH/PH/1$ threshold queue the submatrices $\boldsymbol{R}^{(i)}_{(j,k)}$ are given by:

$$\boldsymbol{R}^{(i)}_{(2,2)} = -\boldsymbol{F}^{(i)}_{(2,2)} \left[\boldsymbol{L}^{(i+1)}_{(2,2)} + \boldsymbol{R}^{(i+1)}_{(2,2)} \boldsymbol{B}^{(i+2)}_{(2,2)}\right]^{-1}, \qquad L \leq i,$$

$$\begin{aligned}
\boldsymbol{R}^{(i)}_{(1,1)} &= -\boldsymbol{F}^{(i)}_{(1,1)} \left[\boldsymbol{L}^{(i+1)}_{(1,1)} + \boldsymbol{R}^{(i+1)}_{(1,1)} \boldsymbol{B}^{(i+2)}_{(1,1)}\right]^{-1}, && 0 \leq i < U-1,\ i \neq L-2,\\
\boldsymbol{R}^{(i)}_{(1,1)} &= -\boldsymbol{F}^{(i)}_{(1,1)} \left[\boldsymbol{L}^{(i+1)}_{(1,1)} + \boldsymbol{R}^{(i+1)}_{(1,1)} \boldsymbol{B}^{(i+2)}_{(1,1)} + \boldsymbol{R}^{(i+1)}_{(1,2)} \boldsymbol{B}^{(i+2)}_{(1,2)}\right]^{-1}, && i = L-2,\\
\boldsymbol{R}^{(i)}_{(1,1)} &= -\boldsymbol{F}^{(i)}_{(1,1)} \left[\boldsymbol{L}^{(i+1)}_{(1,1)}\right]^{-1}, && i = U-1,
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{R}^{(i)}_{(1,2)} &= -\left[\boldsymbol{R}^{(i)}_{(1,1)} \boldsymbol{R}^{(i+1)}_{(1,2)} \boldsymbol{B}^{(i+2)}_{(2,2)}\right] \left[\boldsymbol{L}^{(i+1)}_{(2,2)} + \boldsymbol{R}^{(i+1)}_{(2,2)} \boldsymbol{B}^{(i+2)}_{(2,2)}\right]^{-1}, && L-1 \leq i < U,\\
\boldsymbol{R}^{(i)}_{(1,2)} &= -\boldsymbol{F}^{(i)}_{(1,2)} \left[\boldsymbol{L}^{(i+1)}_{(2,2)} + \boldsymbol{R}^{(i+1)}_{(2,2)} \boldsymbol{B}^{(i+2)}_{(2,2)}\right]^{-1}, && i = U,
\end{aligned}$$

$$\boldsymbol{R}^{(i)}_{(j,k)} = \boldsymbol{0}, \qquad \text{otherwise.}$$

Finally, we obtain the mean sojourn time by applying Little's Law on the mean queue length:

$$\mathbb{E}[S] = \left[-\boldsymbol{\lambda}\boldsymbol{\Lambda}^{-1}\boldsymbol{e}_p\right] \left[\sum_{n=0}^{\infty} n\boldsymbol{\pi}_{\boldsymbol{n}}\boldsymbol{e}\right].$$

## Acknowledgement

This research is supported by the Centre for Telematics and Information Technology (CTIT) of the University of Twente.

## References

[1] K.I. Ahmed, M.E. Ben-Akiva, H.N. Koutsopoulos, and R.G. Mishalani. Models of freeway lane changing and gap acceptence behavior. In *Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, 1996.

[2] N. Baer, R.J. Boucherie, and J.C.W. van Ommeren. The $PH/PH/1$ multi-threshold queue. Technical report, University of Twente, 2012.

[3] M. Baykal-Gürsoy, W. Xiao, and K. Ozbay. Modeling traffic flow interrupted by incidents. *European Journal of Operational Research*, 195:127–138, 2009.

[4] N. Bellomo and C. Dogbe. On the modeling of traffic and crowds: A survey of models, speculations, and perspectives. *SIAM Review*, 53(3):409–463, 2011.

[5] M Boon. *Polling Models - From theory to traffic intersections.* PhD thesis, Eindhoven University of Technology, 2011.

[6] M. Brackstone and M. McDonald. Car-following: A historical review. *Transportation Research - Part F*, 2:181–196, 1999.

[7] L.W. Bright and P.G. Taylor. Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. *Communications in Statistics - Stochastic Models*, 11(3):497–525, 1995.

[8] D.J. Buckley. A semi-poisson model of traffic flow. *Transportation Science*, 2(2):107–133, 1968.

[9] D. Chowdhury, L. Santen, and A. Schadschneider. Statistical physics of vehicular traffic and some related systems. *Physics Report*, 329:199–329, 2000.

[10] F.R.B. Cruz and J.M. Smith. Approximate analysis of $M/G/c/c$ state-dependent queueing networks. *Computers & Operations Research*, 34:2332–2344, 2007.

[11] F.R.B. Cruz, J.M. Smith, and R.O. Medeiros. An $M/G/c/c$ state-dependent network simulation model. *Computers & Operations Research*, 32:919–941, 2005.

[12] Chandler R.E. E.W., Herman R., and Montroll. Traffic dynamics: Studies in car following. *Operations Research*, 6(2):165–184, 1958.

[13] B.D. Greenshields. A study of traffic capacity. In *Proceedings of the 14th Annual Meeting of the Highway Research Board*, 1935.

[14] D. Heidemann. A queueing theory approach to speed-flow-density relationships. In *Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, 1996.

[15] D. Heidemann. Non-stationary traffic flow from a queueing theory viewpoint. In *Proceedings of the 14th International Symposium on Transportation and Traffic Theory*, 1999.

[16] D. Heidemann. A queueing theory model of nonstationary traffic flow. *Transportation Science*, 35(4):405–412, 2001.

[17] D. Heidemann. Mathematical analysis of non-stationary queues and waiting times in traffic flow with particular consideration of the coordinate transformation technique. In *Proceedings of the 15th International Symposium on Transportation and Traffic Theory*, 2002.

[18] D. Helbing. Traffic and related self-driven many-particle systems. *Reviews of Modern Physics*, 73, 2001.

[19] S.P. Hoogendoorn and P.H.L. Bovy. State-of-the-art of vehicular traffic modeling. *J. Syst. Control Eng.*, 215:283–303, 2001.

[20] A. Hordijk and R. Schassberger. Weak convergence for generalized semi-markov processes. *Stochastic Processes and their Applications*, 12:271–291, 1982.

[21] R. Jain and J.M. Smith. Modeling vehicular traffic flow using $M/G/c/c$ state dependent queueing models. *Transportation Science*, 31(4):324–336, 1997.

[22] G. Latouche and V. Ramaswami. A logarithmic reduction algorithm for quasi-birth-death processes. *Journal of Applied Probability*, 30:650–674, 1993.

[23] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modelling*. ASA-SIAM Series on Statistics and Applied Probability. SIAM, Philadelphia, PA., 1999.

[24] L.M. Le Ny and B. Tuffin. A simple analysis of heterogeneous multi-server threshold queues with hysteresis. In *Proceedings of the Applied Telecommunication Symposium*, 2002.

[25] M.J. Lighthill and G.B. Whitham. On kinematic waves. I. Flood movement in long rivers. *Proceedings of the Royal Society of London, Part A*, 229(1178):281–316, 1955.

[26] M.J. Lighthill and G.B. Whitham. On kinematic waves. II. A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London, Part A*, 229(1178):317–345, 1955.

[27] K. Nagel. Particle hopping models and traffic flow theory. *Physical Review E*, 53(5):4655–4672, 1996.

[28] M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach*. Dover Publications, Inc., New York, 1981.

[29] G.F. Newell. A simplified theory of kinematic waves in highway traffic, Part I: General theory. *Transportation Research - Part B*, 27(4):281–287, 1993.

[30] G.F. Newell. A simplified theory of kinematic waves in highway traffic, Part II: Queueing at freeway bottlenecks. *Transportation Research - Part B*, 27(4):289–303, 1993.

[31] G.F. Newell. A simplified theory of kinematic waves in highway traffic, Part III: Multi-destination flows. *Transportation Research - Part B*, 27(4):305–313, 1993.

[32] D. Ni. Multiscale modeling of traffic flow. *Mathematica Aeterna*, 1(1):27–54, 2011.

[33] H.J. Payne. *Models of freeway traffic and control*. Mathematical Models of Public Systems. Simulation Councils, Inc., 1971.

[34] I. Prigogine and F.C. Andrews. A Boltzmann-like approach for traffic flow. *Operations Research*, 8(6):789–797, 1960.

[35] I. Prigogine and R. Herman. *Kinetic Theory of Vehicular Traffic*. American Elsevier, New York, 1971.

[36] P.I. Richards. Shock waves on the highway. *Operations Research*, 4(1):42–51, 1956.

[37] Y. Sugiyama, M. Fukui, M. Kikuchi, K Hasebe, A. Nakayama, K. Nishinari, S. Tadaki, and S. Yukawa. Traffic jams without bottlenecks - experimental evidence for the physical mechanism of the formation of a jam. *New Journal of Physics*, 10(3), 2008.

[38] T. van Woensel. *Models for uninterrupted traffic flows - A queueing approach*. Uninterrupted traffic flows, Universiteit Antwerpen, 2003.

[39] T. van Woensel and N. Vandaele. Empirical validation of a queueing approach to uninterrupted traffic flows. *4OR*, 4:59–72, 2006.

[40] T. van Woensel and N. Vandaele. Modeling traffic flows with queueing models: A review. *Asia-Pacific Journal of Operational Research*, 24(4):435–461, 2007.

[41] T. van Woensel, B. Wuyts, and N. Vandaele. Validating state-dependent queueing models for uninterrupted traffic flows using simulation. *4OR*, 4:159–174, 2006.

[42] N. Vandaele, T. van Woensel, and A. Verbruggen. A queueing based traffic flow model. *Transportation Research - Part D*, 5(2):121–135, 2000.

[43] R.W. Wolff. *Stochastic Modelling and the Theory of Queues*. Prentice-Hall, Englewood Cliffs, New Jersey, 1989.

[44] S.J. Yuhaski Jr. and J.M. Smith. Modeling circulation systems in buildings using state dependent queueing models. *Queueing Systems*, 4:319–338, 1989.