Bounding queuing system performance with variational theory

Jia Li \textsuperscript{a}, H. Michael Zhang \textsuperscript{b,\ast}

\textsuperscript{a}Center for Transportation Research, University of Texas at Austin, TX 78701
\textsuperscript{b}Department of Civil and Environmental Engineering, University of California at Davis, CA 95616

Abstract

Queuing models are often used for traffic analysis, but analytical results concerning a system of queues are rare, thanks to the interdependence between queues. In this paper, we present an analysis of queuing systems to obtain bounds of their performance without studying the details of individual queues. Queuing dynamics is formulated in continuous-time, subject to variations of demands and bottleneck capacities. Our analysis develops new techniques built on the closed-form solution to a generalized queuing model for a single bottleneck. Taking advantage of its variational structure, we derive the upper and lower bounds for the total queue length in a tandem bottleneck system and discuss its implication for the kinematic wave counterpart. Numerical experiments are conducted to demonstrate the appropriateness of the derived upper and lower bounds as approximations in a stochastic setting.

Keywords: queuing systems; traffic flow; variational solution

1. Introduction

Understanding and quantifying the interplay between traffic and bottlenecks is a central focus of traffic flow theory, and numerous analytical and computational models were developed for this purpose. These models capture different level of details and adopt different mathematical representations. Two most widely used macroscopic continuous-time traffic flow models are the (first-order) kinematic wave model (Lighthill and Whitham, 1955; Richards, 1956) and the point queue model (Vickrey, 1969; Nie and Zhang, 2002; Shen and Zhang, 2008), etc. Conventionally, hyperbolic conservation equations (Dafermos, 2005; Toro, 2009) and ordinary differential equations are respectively the primary mathematical tool to analyze the former and latter. Discrete time formulations of corresponding dynamics (e.g., Daganzo (1994); Lebacque (1996)) are derivable from continuous-time models, using numerical tools such as the finite volume approximations (LeVeque, 1992). In this way, we get consistent models in analytical and numerical domains. Alternative formulations of these models exist, including the Lagrangian models and variational models. In general, these models were developed to study link-level flow dynamics.

System-level traffic queuing dynamics and properties are of particular interests in many regards, especially in the context of traffic control and operations. For instance, a better understanding of gridlock mechanism allows more effective control of arterial traffic in urban areas, which triggers various gating and decentralized strategies in recent
years (see e.g. Keyvan-Ekbatani et al. (2012) and Li and Zhang (2014) and references therein). As another example, in operations of a corridor, it is desirable to consider collective dynamics of tandem queues along sequential bottlenecks rather than addressing them isolatedly, due to the their underlying interdependence. Nonetheless, albeit the desirability to consider queuing systems as an entity, knowledge on analytical properties of such systems are limited and difficult to obtain, which constitutes a major hurdle to developing a sound systemic theory and related applications. Among others, a central problem to conquer in this agenda is how to establish aggregate queuing properties based on given queuing dynamics at link and node level. In the last several years, we have seen increasing studies along this direction, which attempted to build the linkage between local (link- and node-level) queuing dynamics and analytical properties of networks. The properties that have been investigated include existence of macroscopic fundamental diagram in traffic networks (Daganzo and Geroliminis, 2008), existence of solution to a dynamical queuing system Jin (2012), stability of diverge-merge networks (Jin, 2013), and continuity of path delay operators on general networks (Han and Friesz, 2012; Han et al., 2015). In these studies, link models employed range from simple link queue model to the more full-fledged LWR model, but their focuses are exclusively on qualitative characters. Quantitative measures, such as level of service (LoS), throughput of network, etc., were not addressed. Along another line of research, which tackles the so-called morning commute problem, exact arrival profiles are derivable, thus giving full quantitative characterizations of the system (see e.g. Newell (1988) and Kuwahara (1990)). Nonetheless, similar analysis can become demanding and difficult to extend if more than two bottlenecks are involved.

We aim to bridge this gap. As a first step, in this paper, we consider the problem of performance bounding. This problem looks for upper and lower bounds of system performance, when demand and bottleneck capacity data of a queuing system are given. In other systems, e.g. communication networks, this problem is well known, and the network calculus theory (Cruz, 1991; Le Boudec and Thiran, 2001) was developed to tackle it. Some models in the network calculus, e.g. the leaky bucket model, are similar to the point queue model that we discuss below. But as will be clear from below, the approach developed in this paper is different from network calculus. Central to the network calculus are the concepts of service curve and max-plus algebra, based on which link-level dynamics are concatenated. In contrast, our analysis exploits a variational property of point queue dynamics in this paper. Performance bounds are established through making use of this special structure. Moreover, relations between the point queue model and kinematic wave model are examined, in order to extend the analysis concerning point bottleneck models.

We take the following steps. We start from a simple yet flexible queuing model, called generalized queuing model (Li and Zhang, 2015). We present this model and review its major properties. Among others, the variational property of this model is most interesting and relevant, which allows its solution be expressed in closed form, even when demand and bottleneck capacity are time-dependent and discontinuous. Furthermore, the closed form of this solution is mapping of nothing else but the so-called demand surplus function \( h(t) \). Upon noting this link-level property, we demonstrate that the notion of demand surplus can be extended to a more general setting, i.e. a route consisting of tandem bottlenecks. Such an extension underpins upper and lower bounds for system performance, e.g. total queue length. Having bounded the performance of a general tandem queue system analytically, we consider approximation of first-order kinematic wave model with the generalized queuing model, in order to understand the impact of different notions of queue on resulted performance bounding. We derive their difference, in terms of queue length, when identical initial and boundary data are used. Along with these steps, we also prove the tightness of given bounds, and conduct numerical experiments to verify our analytical findings.

The remainder of this paper is organized as follows. In Section 2, we introduce the technical background of this study. In Section 3, we present the major result of this paper, performance bounds derived from the generalized queuing model, extension of the results to kinematic wave models. In Section 4, we conduct numerical experiments and verify derived analytical results. In Section 5, a discussion is presented concerning the modeling issues arising in previous sections. In Section 6, we summarize the findings and discuss future works.

2. Preliminaries

In this section, we provide an overview of queuing models, queuing system analysis, and the variational theory of traffic flow. We start with considering traffic dynamics on a link. The following notation will be used:

\[ t: t \in \mathbb{R}^+, \text{ time} \]
Queuing models captures dynamic demand-supply interplay in various service systems, e.g. transportation network, communication network, and logistic network. Following are some widely used models in traffic flow, ordered by complexity.

**Delay function model.** Delay function models are widely used in the traffic assignment literature (e.g. Friesz et al. (1993)). They say that the link traversal time is a function of queue length at time $t$, i.e.

$$\tau(t) = f(x(t)) \quad (1)$$

To be logically coherent, the FIFO property is desirable, and this requires that $f(\cdot)$ is linear (Nie and Zhang, 2002).

**Point queue model.** This model was proposed in Vickrey (1969). Queue evolution is modeled by outflow dynamics (3) in conjunction with a conservation equation (2):

$$\dot{x}(t) = u(t) - v(t), \quad t \geq 0 \quad (2)$$

and

$$v(t) = \begin{cases} 
C & x(t) > 0 \\
\min(C, u(t - \tau_0)) & x(t) = 0 
\end{cases} \quad (3)$$

**Spatial queue model.** Based on the point queue model, spatial queue model captures spillover issue modifying outflow dynamics, reflecting the influence of downstream storage capacity $S_d$ (Nie, 2003). In the case of two tandem bottlenecks (downstream link is labeled by $d$), the outflow from the first bottleneck is

$$v(t) = \begin{cases} 
C & x(t) > 0, x_d(t) < S_d \\
u(t - \tau_0) & x(t) = 0, x_d(t) < S_d \\
0 & x_d(t) = S_d 
\end{cases} \quad (4)$$

**LWR model.** The LWR model is a conservation law with a flow-density fundamental diagram that captures the driver’s tendency to slow and as vehicle density increases on the road. It models the spatiotemporal evolution of traffic density $\rho(x, t)$ on a link. The model reads,

$$\partial_t \rho(x, t) + \partial_x Q_\rho(\rho(x, t)) = 0, \quad (x, t) \subset (-\infty, \infty) \times [0, \infty) \quad (5)$$

where $Q_\rho(\cdot)$ is fundamental diagram. The link inflow and outflow are dictated by boundary conditions, which could be a merge model or a diverge model or a static bottleneck.

Almost all continuous-time traffic queuing models are connected to the above models in some way, either with varying level of details or in alternative formulations. A thorough relevant discussion is presented in Jin (2014). We use two examples to illustrate. The first example is the model of Bliemer (2007). This model considers a point bottleneck, but capturing more details by decomposing traffic over the link into free flowing and queuing part, and capturing queue spillover through a node model. Another example is that, as noted in Astarita (2002), given link traversal time function $\tau(t)$, one can derive its equivalent outflow dynamics as follows,

$$v(t + \tau(t)) = \frac{u(t)}{1 + d\tau(t)/dt} \quad (6)$$
2.2. Queuing system analysis

In general, a queuing system is defined on a directed graph \(G(\mathcal{N}, \mathcal{A})\), where \(\mathcal{N}\) and \(\mathcal{A}\) respectively denote the set of nodes and arcs. As such, the system performance is determined by two factors, namely the dynamics over \(\mathcal{A}\) as well as interplay of the dynamics through \(\mathcal{N}\). The interplay through \(\mathcal{N}\) determines how aggregate dynamics can be derived, and in this point of view, queuing models can be categorized according to whether queue spillover is addressed. Difference of these two types of model are intuitive: when queue spillover does not exist, effect of queuing is confined locally in space, and will not spread out over the network. In contrast, gridlock may be resulted from the network-wide spatial interactions of queues when spillover occurs. A numerical investigation of this difference is presented in Zhang et al. (2013), and the results indicate when spillover is ignored, travel time cost can be underestimated.

Markov chain and network calculus are two primary tools to analyze systems consisting of point bottlenecks. The Markov chain is usually used to study stability (boundedness of queues in the long run) of discrete-time queuing networks, where boundary inputs are stochastic processes. Interested readers may refer to Tassiulas and Ephremides (1992), Giaccone et al. (2005) and Varaiya (2013). In contrast, network calculus (Cruz, 1991; Le Boudec and Thiran, 2001) is used to calculate bounds on performance measure such as delay and queue length. Similar to traditional queuing theory (Newell, 1982) and the variational theory of traffic flow (Newell, 1993; Daganzo, 2005), network calculus considers relations between cumulative flows. Network calculus is centered on the concept of service curve \(S(\cdot)\) (Fidler, 2010). A wide sense increasing function \(S(\cdot)\) is a service curve if and only if \(S(0) = 0\) and \(V(t) \geq \inf_{0 \leq s \leq t} [U(s) + S(t-s)], \quad \forall t \geq 0\). This definition per se gives the lower bound of cumulative outflow at a point bottleneck (when \(t_0 = 0\)). The definition can also be reformulated as: \(S(\cdot)\) is a service curve if and only if \(\forall t \geq 0, 3s \in [0, t]\), such that \(V(t) \geq U(s) + S(t-s)\), equivalently, \(V(t) - V(s) \geq U(s) - V(s) + S(t-s) = \alpha(s) + S(t-s)\). Based on the definition, it is straightforward to derive a tight upper bound on queue length: \(x(t) \leq \sup_{s \leq t} [\alpha(s) - S(s)]\), where \(\alpha(t)\) is the arrival curve that bounds \(U(t), U(t) - U(s) \leq \alpha(t-s), \forall s \leq t\). A new max-plus algebra can be defined on space of increasing functions \(\mathcal{F}_0\), which equipped with operations \(\otimes\): \(\mathcal{F}_0 = \{g : g(t) \geq g(s) \geq 0, \quad \forall t \geq s, g(0) = 0\}\) and \(g_1 \otimes g_2(t) = \inf_{0 \leq s \leq t} [g_1(s) + g_2(t-s)]\). Based on the new algebra, service curve for a tandem queue system can be derived through concatenating individual ones: Consider a tandem bottleneck system, whose bottlenecks have service curve \(S_1, \ldots, S_n\). Then the tandem system has the following service curve, expressed in the max-plus algebra:

\[
S_{net} = S_1 \otimes S_2 \otimes \ldots \otimes S_n(t) \quad (7)
\]

With this service curve, the performance bounds for a system can be obtained. This philosophy is pursued in the following, but from a distinct approach. Instead of using the service curve, we exploit the variational solution to a generalized queuing model and repeatedly use the notion of demand surplus \(h(t)\) associated with it. System behaviors are characterized through aggregating queuing dynamics at individual bottlenecks.

3. Bounding queuing system performance

3.1. Problem statement

In a broad context, we consider a system consisting of interconnected bottlenecks, which forms a directed acyclic graph as illustrated in Figure 1. There are three types of nodes, i.e. entry nodes (\(E1, E2\)), exit nodes (\(X1\) to \(X3\)), and ordinary nodes (\(O1, O2\)), whose meaning are clear from the figure. The problem is: given the demands to the entry nodes and capacity of each bottleneck, as well as necessary routing information (we assume it is exogenous in this paper), estimate the performance of this queuing system, in terms of e.g. upper and lower bounds of system queuing delay, throughput, and total queue length. We assume demand and capacity are dynamic over continuous-time, and demand and capacity are exogenous and independent. Implication of this problem is evident: suppose we know the capacity values at individual bottlenecks and routing information (e.g. turning ratio), how can we provide upper and lower bound estimates on the performance of this system.

As the first step to tackle the general problem, in this paper we demonstrate how this problem is tackled when tandem bottleneck (queue) systems are concerned. We take the following steps. We start with introducing a generalized point queue model, whose capacity is \(C(t)\), i.e. time-dependent. For this model, we can derive its closed-form solution and thus know its exact performance (queue length, queuing delay, etc.). Then we analyze the aggregate dynamics of
Fig. 1. Illustration of the problem: flow through connected bottlenecks

a tandem queue system through analyzing the interdependence of its components. Variational property of the generalized queuing dynamics play a critical role in this step. We obtain upper and lower bounds on total queue length, and demonstrate that they are tight. At last, we discuss the approximate relation between the generalized queuing model and the LWR model, when the same boundary conditions are imposed. Following this section, numerical experiments are conducted to verify the analytical results in this Section.

3.2. Generalized point queue model and its variational solution

The variational theory, in the most general sense, refers to the theory exploiting variational principle of a dynamical process, i.e.

\[ \delta \int \int L \, dx \, dt = 0 (8) \]

where \( L \) is the Lagrangian function. Mathematical analysis of a dynamic process that is otherwise difficult is usually simple with this formulation. The study of variational principles stem from the field of mechanics (Lanczos, 1970) and fluid mechanics (Whitham, 1967). In kinematic wave models, it often refers to application of Lax-Hopf formula which solves the Hamilton-Jacobi equation. In particular, the cumulative traffic count \( N(x,t) \) of the following form solve the Hamilton-Jacobi equation,

\[ N(x,t) = \inf_{B \in \mathcal{B}, \gamma \in \mathcal{P}} \left\{ N_B + \int_{\gamma} \{ Q_\epsilon (\rho(s), s) - \rho \gamma(s) \} ds \right\} (9) \]

where \( \mathcal{B} \) and \( \mathcal{P} \) are respectively the set of boundary points where initial-boundary data are given and set of paths connecting the boundary and point \((x,t)\). It can be proved that the infimum on the right hand of (9) can only be attained along wave paths, and the optimum path (denoted as \( \gamma^*(s), s \in \{0,t]\) ), as suggested by the form of (9), solves an optimal terminal cost problem.

Interestingly, though distinct from the LWR model in formulation (i.e. ODE (ordinary differential equation) vs. PDE (partial differential equation)), Vickrey’s model also admits a variational solution. Without loss of generality, let
\[ \tau_0 = 0. \] Then (2) and (3) admit the following solution,

\[ V(t) = Ct + \inf_{s \leq t} (U(s) - Cs), \quad t \geq 0 \tag{10} \]

which is by formulating the Vickrey’s model into a Hamilton-Jacobi equation (Han et al., 2013). In Li and Zhang (2015), (10) is further generalized, removing the constraint on function form of bottleneck capacity. This is achieved through measure theoretic analysis with the life cycle of queuing process. In particular, consider the point queue model equipped with the following outflow dynamics:

\[ v(t) = \begin{cases} 
C(t) & \text{if } x(t) > 0 \\
\min\{C(t), u(t - \tau_0)\} & \text{otherwise} 
\end{cases} \tag{11} \]

The main idea to derive corresponding exact solution is stated here. First, associated with each queuing process, regardless of details of inflow and capacity profiles, the link switches between two and only two states, i.e. when a queue exists on the link (i.e. \( x(t) \geq 0 \) for \( t \) in some interval), and there is no queue. In the former scenario, during the time interval when a queue exists, the queue experiences growth, decay and diminish. We call this process a life cycle of the concerned queue, and label queue with integer \( i \). During a complete life cycle of a queue, a conservation property must hold, still independent of the inflow and capacity profile details (but of course, the cumulative amount), i.e.: cumulative flow discharged from this queue equals to the cumulative flow joining the queue. Moreover, the discharged flow is equal to the capacity value \( C(t) \) by the definition of outflow. Therefore, we have

\[ \int_{O_i} (u(s) - C(s))ds = 0 \tag{12} \]

where \( O_i \) is the life span of the \( i \)-th queue. When there is no queue, let’s denote the time between \( O_i \) and \( Q_{i+1} \) as \( \bar{O}_i \). Then over \( \bar{O}_i \), there is

\[ \int_{\bar{O}_i} (u(s) - C(s))ds < 0 \tag{13} \]

since discharging flow is no more than capacity on any set of positive measures. These observations lead to the consideration of demand surplus in the following theorem, which can be decomposed accordingly. It turns out queue length of this generalized queuing process has a succinct form, which is the deviation between \( h(t) \) and its infimum over \([0, t)\). Theorem 1 states this results, whose detailed proof is provided in Li and Zhang (2015).

**Theorem 1** (Variational solution to the generalized queuing model). *Without loss of generality, assume \( \tau_0 = 0 \), the generalized queuing model admits the following solution,

\[ x(t) = h(t) - \inf_{s \leq t} h(s) \tag{14} \]

where

\[ h(t) = \int_{(0,t]} (u(s) - C(s))ds \tag{15} \]

is called demand surplus. The solution formula holds when \( C(t) \) is piecewise Lipschitz continuous.*

Some observations & remarks:

1. Demand surplus \( h(\cdot) \) represents the difference between cumulative demand and supply at a bottleneck. The interesting implication of Theorem 1 is that the queue length at any instant \( t \) is simply a function of demand surplus. This agrees with the conventional queuing theory, where the difference between instantaneous inflow and bottleneck capacity determines queue evolution.

2. Compared to the conventional queuing theory, the major advantage of (14) lies in the closed form expression that captures overall queuing dynamics. This allows evaluation of queue lengths at any instant \( t \) using only boundary data (i.e. capacity and inflow values), without considering/tracking intermediate details of the state (i.e. queue length before \( t \)).
3. Traversal time through a bottleneck consists of two parts: queuing delay and free flow travel time. Assuming $\tau_0 = 0$ implies that the free flow travel time is independent of the queuing delay. This is true when the queue is ‘vertical’, i.e. does not have physical lengths. Relaxation of this assumption will follow.

4. The relaxation of bottleneck capacity to a generic function $C(t)$ allows to capture some significant effects, which include, among others, traffic signal control and queue spillover. In the latter case, $C(t)$ is a function of downstream queue lengths and storage capacities.

5. Though intuitive, well-posedness of the generalized queuing model is not obvious, since its right-hand side can be discontinuous, thus resulting in a non-smooth ODE (see e.g. the discussion in Ban et al. (2012)). We provide relevant proof in Li and Zhang (2015). Here we give a proof on its FIFO property. See Proposition 1.

**Proposition 1** (FIFO property). The generalized queuing model (equations (2) and (14)) is FIFO.

**Proof.** Given $t \geq 0$, we consider $t' > t > 0$, where $|t' - t| << 1$. Denote the travel time (i.e. queuing delay when $\tau_0 = 0$) of traffic entering at $t$ as $\tau(t)$, it is evident that $\tau(t)$ should be solved as

$$
\tau(t) = \inf_{b \geq 0} \left\{ \int_{[t,t+b]} C(s)ds \geq x(t) \right\}
$$

from the equation

$$
\int_{[t,t+\tau(t)]} C(s)ds = x(t)
$$

The same equation also holds for $t'$, i.e.

$$
\int_{[t',t'+\tau(t')]} C(s)ds = x(t')
$$

Subtracting (18) from (17) and doing some algebra, we have

$$
\int_{[t'+\tau(t'),t+\tau(t)]} C(s)ds = x(t) - x(t') - \int_{[t,t']} C(s)ds = \int_{[t,t']} (u(s) - C(s))ds - \int_{[t,t']} C(s)ds = \int_{[t',t]} u(s)ds
$$

Since $u(\cdot) \geq 0$, $C(\cdot) \geq 0$ and $t' > t$ as we assumed, excluding the trivial cases (i.e. $u(s) = 0$ or $C(s) = 0$ a.e. $\forall s \in [t,t']$), there must be $t' + \tau(t') > t + \tau(t)$, i.e. the model is FIFO.

To conclude this section, we point out that upon getting the exact solution (14), which is queue length, other performance measures such as queuing delay and link traversal time are immediately obtainable.

### 3.3. Performance bounds for tandem bottlenecks

Naturally, the next step is considering flow traversing tandem bottlenecks. Motivated by (7), where the system service curve is obtained through concatenating individual service curves, we extend the notion of demand surplus from a single bottleneck to tandem bottlenecks. The main result is Theorem 2.

**Theorem 2** (Performance bounds for tandem bottlenecks). In a tandem queuing system with $N$ bottlenecks (labeled as $1, \ldots, N$ from upstream to downstream), the total length of queues satisfies the following inequality,

$$
\overline{h}_N(t) - \inf_{r \leq t} \overline{h}_N(r) \leq \sum_{n=1}^{N} x_n(t) \leq \underline{h}_N(t) - \inf_{r \leq t} \underline{h}_N(r), \ \forall t \geq 0, N \geq 1
$$

where

$$
\underline{h}_N(t) = \int_{[0,t]} (u(s) - \min_{1 \leq n \leq N} C_n(s))ds
$$

and

$$
\overline{h}_N(t) = \int_{[0,t]} (u(s) - C_N(s))ds
$$

with $u(t)$ being the arrival flow to the system.
We prove the lower bound first. In case $N = 1$, the inequality holds due to Theorem 1, and actually the equality is always attained. We consider the case $N = 2$. Denote the departure flow from bottleneck 1 as $v_1(t)$. It is meanwhile the arrival flow of the bottleneck 2. We can evaluate $h_2(t)$,

$$h_2(t) = \int_0^t (v_1(s) - C_2(s))ds$$

$$= \int_0^t (C_1(s) - C_2(s))ds + \inf_{r \leq t} \int_0^r (u(s) - C_1(s))ds$$

$$= \int_0^t (C_1(s) - C_2(s))ds + \inf_{r \leq t} h_1(r) \tag{23}$$

Therefore, we have

$$x_1(t) + x_2(t) = h_1(t) - \inf_{r \leq t} h_1(r) + h_2(t) - \inf_{r \leq t} h_2(r)$$

$$= h_1(t) + \int_0^t (C_1(s) - C_2(s))ds - \inf_{r \leq t} (\int_0^r (C_1(s) - C_2(s))ds + \inf_{s \leq r} h_1(s)) \tag{24}$$

Note that

$$\int_0^t (C_1(s) - C_2(s))ds + \inf_{r \leq t} h_1(s) \leq \int_0^t (C_1(s) - C_2(s))ds + h_1(t) = \int_0^t (u(s) - C_2(s))ds \tag{25}$$

We have

$$x_1(t) + x_2(t) \geq \int_0^t (u(s) - C_2(s))ds - \inf_{r \leq t} \int_0^r (u(s) - C_2(s))ds = \tilde{h}_2(t) - \inf_{r \leq t} \tilde{h}_2(r) \tag{26}$$

To generalize this result for the case $N > 2$, we simply note that

$$h_{N+1}(t) = \int_0^t (C_N(s) - C_{N+1}(s))ds + \inf_{r \leq t} h_N(r) \tag{27}$$

and for $N \geq 1$

$$h_{N+1}(t) \leq \int_0^t (C_N(s) - C_{N+1}(s))ds + h_N(t) \tag{28}$$

Using (28) iteratively, we have

$$h_{N+1}(t) \leq \int_0^t (u(s) - C_{N+1}(s))ds \tag{29}$$

Then if the inequality holds for $N$, it also holds for $N + 1$, because

$$\sum_{n=1}^{N+1} x_n(t) = x_{N+1}(t) + \sum_{n=1}^N x_n(t)$$

$$\geq h_{N+1}(t) - \inf_{r \leq t} h_{N+1}(r) + \tilde{h}_N(t) - \inf_{r \leq t} \tilde{h}_N(t)$$

$$= \int_0^t (C_N(s) - C_{N+1}(s))ds + \tilde{h}_N(t) - \inf_{r \leq t} \tilde{h}_N(t) \tag{30}$$

We thus complete the proof for the first inequality, i.e. the lower bound of $\sum_{n=1}^N x_n(t)$, by induction.

Now we look into the upper bound of $\sum_{n=1}^N x_n(t)$. In case $N = 1$ the inequality obviously holds, and actually the equality is attained. Note in general, the total queue length is upper bounded by a worst case scenario when all bottlenecks have the same capacity as the most stringent bottleneck. This proves the upper bound for $\sum_{n=1}^N x_n(t)$. ⧫

The following property of tandem bottleneck is intuitive but worth mentioning, which helps to understand the temporal and spatial impacts of capacity variations on queuing.

**Lemma 1** (Monotonicity). In a tandem queue system, consider the mapping $\phi : (u(t), C_1(t), \ldots, C_N(t)) \mapsto \sum_{n=1}^N x_n(t)$. The following monotone property holds: if $C_n(t) \leq \bar{C}_n(t)$ for an arbitrary $n$ ($1 \leq n \leq N$), then

$$\phi(u(t), C_1(t), \ldots, C_n(t), \ldots, C_N(t)) \geq \phi(u(t), C_1(t), \ldots, \bar{C}_n(t), \ldots, C_N(t)) \tag{31}$$

In general, this property holds when $(C_n(t))_{1 \leq n \leq N} \leq (\bar{C}_n(t))_{1 \leq n \leq N}$ in the component-wise sense.
First consider the case \( N = 1 \). Let \( \tilde{C}(t) = C(t) + \delta C(t) \), where \( \delta C(t) \geq 0 \) for all \( t \). It is straightforward to verify, 
\[
\phi(u(t), \tilde{C}(t)) - \phi(u(t), C(t)) = -f(t) - \inf_{s \in [0, t]} (g(s) - f(s)) + \inf_{s \in [0, t]} g(s) \\
\leq -f(t) - (\inf_{s \in [0, t]} g(s) + \inf_{s \in [0, t]} (-f(s))) + \inf_{s \in [0, t]} g(s) \\
= -f(t) + \sup_{s \in [0, t]} f(s) = 0
\]
where \( f(s) = \int_0^s \delta C(l)dl \), \( g(s) = \int_0^s (u(l) - C(l))dl \), and we use a property of infimum in the second line and non-negativity of \( \delta C(t) \) in the third line. When \( N = 2 \), the case \( n = 2 \) is evident, since in the two systems the first queue and corresponding discharge flow are identical and we can apply (32) to the second bottleneck. If \( n = 1 \), (24) implies \( x_1(t) + x_2(t) = \int_0^t (u(s) - C_2(s))ds - \inf_{s \in [0, t]} \left( \int_0^s (C_1(s) - C_2(s))ds + \inf_{s \in [0, t]} \int_0^s (u(l) - C_1(l))dl \right) \). So it suffices to evaluate the monotonicity of mapping \( \mu(C_1) \equiv \int_0^t (C_1(s) - C_2(s))ds + \inf_{s \in [0, t]} \int_0^s (u(l) - C_1(l))dl \) with respect to \( C_1(t) \). It is straightforward to verify, 
\[
\mu(C_1 + \delta C_1) = \int_0^t (C_1(s) - C_2(s))ds + \int_0^t \delta C_1(s)ds + \inf_{s \in [0, t]} \left( \int_0^s (u(l) - C_1(l))dl - \int_0^s \delta C_1(l)dl \right) \\
= \int_0^t (C_1(s) - C_2(s))ds + \inf_{s \in [0, t]} \left( \int_0^s (u(l) - C_1(l))dl + \int_0^s \delta C_1(l)dl \right) \\
\geq \mu(C_1)
\]
where the second line is due to the non-negativity of \( \delta C_1(t) \). Therefore \( \phi \) decreases when \( C_1 \) increases, which is exactly what we want to prove. In general, suppose the result holds for \( N - 1 \), then in the case \( N \), the only non-trivial case to consider is \( n = 1 \), because otherwise it reduces to the case of \( N \). We note that from (14) and (27), there is
\[
\sum_{n=1}^N x_n(t) = \int_0^t (u(s) - C_N(s))ds - \inf_{s \in [0, t]} \int_0^s (u_N(l) - C_N(l))dl
\]
Based on (14), it can be shown that uniformly larger \( C(t) \) and \( U(t) \) both lead to uniformly larger cumulative outflow \( V(t) \). So in a tandem bottleneck system, \( \delta C(t) > 0 \) results in uniformly larger \( \int_0^t u_N(l)dl \), implying that the total queue length \( \sum_{n=1}^N x_n(t) \) becomes smaller based on (34). In the end, the most general case (i.e. varying multiple \( C_n(t) \) simultaneously) is evidently true, since comparison can be done by introducing intermediate inequalities, which vary one \( C_n(t) \) a time.

Lemma 1 says that increasing any bottleneck(s) capacity in a tandem system improves the overall system performance for all the time. However, it is possible that queuing may become worse at certain bottlenecks at certain time, so the improvement is not uniform in space.

The upper and lower bounds in Theorem 2 are tight. To see this, the upper bound of total queue size is determined by the most stringent bottleneck in the tandem. This bound is reached when e.g. one bottleneck is constantly the most restrictive one, i.e. there exists \( n^* \in \{1, \ldots, N\} \) such that
\[
C_{n^*}(t) = \min_{1 \leq s \leq N} C_s(t), \forall t \geq 0
\]
On the other hand, the lower bound on total queue length is determined by the capacity of the last bottleneck, i.e. \( C_N(t) \). This bound is tight when
\[
N = \arg \min_{1 \leq s \leq N} C_s(t), \forall t \geq 0
\]
i.e. the last bottleneck is the most stringent.

To better understand the system character, let’s introduce the notion of effective capacity, which we denote as \( C_e(t) \). For the tandem system discussed in this section, \( C_e(s) (s \geq 0) \) is a function satisfying
\[
\sum_{n=1}^N x_n(t) = h_e(t) - \inf_{0 \leq s < t} h_e(s)
\]
where \( h_e(t) = \int_0^t (u(s) - C_e(s))ds \). Theorem 2 and Lemma 1 imply the following property about \( C_e(t) \).

Corollary 1 (Effective capacity of tandem bottlenecks). Effective capacity \( C_e(t) \) of a system of tandem bottlenecks satisfies the inequalities
\[
\min_{1 \leq s \leq N} C_s(t) \leq C_e(t) \leq C_N(t)
\]
Proof. Obvious from Theorem 2 and Lemma 1. □

This corollary characterizes the goodness of approximation of the bounds in Theorem 2. It is seen that the given upper and lowers bounds tend to the true value of total queue length, when \( \| \min_{1 \leq n \leq N} C_n(t) - C_N(t) \| \) tends to zero.

3.4. Extension to networks

We consider a set of flows traversing a network consisting of multiple bottlenecks. This setup is similar to Varaiya (2013) within a traffic signal control context, and Tassoulas and Ephremides (1992) and Giaccone et al. (2005) within a packet network context. In these literature, bottleneck capacity allocation (control) policies are sought, so as to achieve certain system-level objective, such as maximal network throughput. Routing policies are prescribed or determined through online estimation.

We label path of a stream as \( p(p = 1, \ldots, P) \), bottlenecks as \( n(n = 1, \ldots, N) \), bottleneck capacity as \( C_n(t) \) \( (t \geq 0) \). The set of bottlenecks along path \( p \) is written as \( B(p) = \{ n_1^p, \ldots, n_p^p \} \), where the subscript aligns with the direction of flow from upstream to downstream and \( |p| \) is the total number of bottlenecks that \( p \) traverses. Similarly, denote the set of paths traversing bottleneck \( n \) as \( P(n) = \{ p_1^n, \ldots, p_{|n|}^n \} \), where \( |n| \) is the total number of paths traversing this bottleneck. The system is acyclic if and only if values in \( P(n) \) are distinct, for all \( p \). Similar to the literature mentioned above, it is assumed that \( |n| \) separate point queues form upstream of the bottleneck \( n \). As such, we have a system of \( \sum_{n=1}^{N} |n| \) queues, with lengths \( x_{np}(t) \), for all \( p \) and \( n \in B(p) \). A control policy is a mechanism of allocating \( C_n(t) \) to these queues, according to e.g. the current system state and/or certain forecasts. Examining detailed control policy design and properties is beyond the scope of this paper. Here we assume information on division of queues, according to e.g. the current system state and/or certain forecasts. Examining detailed control policy design and properties is beyond the scope of this paper. Here we assume information on division of queues, according to e.g. the current system state and/or certain forecasts.

We consider a set of flows traversing a network consisting of multiple bottlenecks. This setup is similar to Varaiya (2013) within a traffic signal control context, and Tassoulas and Ephremides (1992) and Giaccone et al. (2005) within a packet network context. In these literature, bottleneck capacity allocation (control) policies are sought, so as to achieve certain system-level objective, such as maximal network throughput. Routing policies are prescribed or determined through online estimation.

We consider a set of flows traversing a network consisting of multiple bottlenecks. This setup is similar to Varaiya (2013) within a traffic signal control context, and Tassoulas and Ephremides (1992) and Giaccone et al. (2005) within a packet network context. In these literature, bottleneck capacity allocation (control) policies are sought, so as to achieve certain system-level objective, such as maximal network throughput. Routing policies are prescribed or determined through online estimation.

Obvious from Theorem 2 and Lemma 1.

3.5. Approximation of kinematic wave models

Nie and Zhang (2002) observed that point queue models can be an approximation to more realistic kinematic wave models, in the sense that travel time predicted by the two models are consistent. This is not quite surprising. The major difference of the two models lie in the notion of queue. A queue in the LWR model is a shockwave, which takes road space, thus called ‘horizontal queue’. In contrast, the queue we discussed above can be regarded as ‘vertical queues’.
queue", since the physical space it takes is irrelevant to the dynamics we consider. Despite of this difference, it turns out link-level queuing delay is independent of how the queue is defined.

In this section, we generalize the observation of Nie and Zhang (2002), by considering links with time-dependent discharging boundaries. The purpose is to show that the performance bounds derived above can serve as good approximations in more realistic settings under certain circumstances. To be precise, we consider an initially empty link whose downstream boundary is controlled with a time-dependent capacity $C(t)$; link length is $L$; cumulative flow of inflow traffic is $U(t)$. Traffic can be modeled in two ways,

- Generalized queuing model: model is parameterized by $\tau_0$, with data $C(t)$ and $U(t)$.
- LWR model with time-dependent boundary: model is parameterized $L$ and a triangle fundamental diagram parameterized by the tuple $(v_f, -w, C_{\text{max}})$ (i.e. free flow speed, shockwave speed, and capacity). Boundary conditions are given by $C(t)$ and $U(t)$.

To ensure consistency, it is required $v_f = L/\tau_0$, and $0 \leq C(t) \leq C_{\text{max}}$ for all $t \geq 0$. Our main observation is Theorem 3. The basic idea is illustrated in Figure 2. We call a vehicle in queuing regime if its speed is less than $v_f$. Spatiotemporal regions within which vehicles are in the queuing regime can be identified. We call them queues. Then we can group incoming traffic according to whether these vehicles join the same queue. For instance, corresponding to $BCC'B'$, traffic entering between $[t_B, t_C]$ later join and leave from the same queue; corresponding to $CDD'C'$, all entry vehicles traverse the link at free flow speed. For different groups of traffic, like in the generalized queuing model, their demand surplus switches between negative values and zeros. This fact is used to derive the number of queuing vehicle at any instant. See below. The following property is needed.

**Assumption 1.** Consider kinematic waves on a link $(a, b)$ whose downstream discharging capacity is $C(t)$, which is less than or equal to $C_{\text{max}}$ for all $t \geq 0$. Suppose $t_i$ and $t_e$ are respectively the departure time of the first and last vehicle pertaining to a queue, then

$$\int_{t_i}^{t_e} u(s) ds = \int_{t_i + \tau_0}^{t_e + \tau_0} C(s) ds$$

(41)

In another word, $v(s) = C(s)$ a.e. over $(t_i, t_e)$.

Note this assumption can be proved in simple cases, e.g. when $C(s)$ emulates green-red signals, and $u(s)$ is constant. This assumption is also intuitive, which essentially implies that during a queuing life cycle, the discharging flow is
equal to the bottleneck capacity pointwisely. However, in the most general case as in the generalized queuing model, e.g. when \( C(t) \) is allowed to be piecewise continuous, we haven’t established a rigorous proof of (41).

**Theorem 3.** Refer to the LWR model with time-dependent boundary as described above and scenario depicted in Figure 2, when spillover doesn’t occur, there is

\[
n(t_G) = h(t_G) - \inf_{s \leq t_G} h(s) ds + \int_{(0,t_G]} u(s) ds
\]

where \( n(t_G) \) is the number of queuing vehicles at time \( t_G \), \( t_E \) is the entry time of the last queuing vehicle at \( t_G \), and similar to the generalized queuing model

\[
h(t) = \int_{(0,t)} (u(s - \tau_0) - C(s)) ds
\]

**Proof.** We have the following relations (we denote \([t_A, t_B]\) as \( I_{AB} \), \([t_A', t_B']\) as \( I_{AB}'\):

\[
\int_{I_{XY}} u(s) ds = \int_{I_{XY}} v(s) ds, \text{ for } XY = AB, BC, CD
\]

\[
v(t) \leq C(t), \text{ for } t \in I_{AB}' \cap I_{CD}' \cap v(t) = C(t), \text{ for } t \in I_{AB}' \cap I_{CD}'
\]

Correspondingly on \( ABB'A' \) and \( CDD'C' \) we have

\[
\int_{I_{XY}} (u(s - L/v_f) - C(s)) ds < 0, \text{ for } XY = AB, CD
\]

and

\[
\int_{I_{XY}} (u(s - L/v_f) - C(s)) ds = 0, \text{ for } XY = BC
\]

Combined, they imply on \( ADD'A' \)

\[
\inf_{s \leq t_G} \int_{[L/v_f, s]} (u(s - L/v_f) - C(s)) ds = \int_{I_{AB}} (u(s - L/v_f) - C(s)) ds
\]

Applying Green’s theorem to \( DEFGD' \), we have

\[
\int_{[l_{D}, l_{E}]} u(s) ds = n(t_G) + \int_{[l_{D}, l_{G}]} C(s) ds
\]

Therefore

\[
n(t_G) = \int_{[l_{D}, l_{E}]} u(s) ds - \int_{[l_{D}, l_{G}]} C(s) ds
\]

We plug (48) into this relation,

\[
n(t_G) = n(t_G) + 0
\]

\[
= \int_{[l_{D}, l_{E}]} u(s) ds - \int_{[l_{D}, l_{G}]} C(s) ds + \int_{[l_{D}, l_{G}]} (u(s - L/v_f) - C(s)) ds - \inf_{s \leq t_G} \int_{[L/v_f, s]} (u(s - L/v_f) - C(s)) ds
\]

\[
= \int_{[l_{D}, l_{E}]} u(s) ds - \int_{[l_{D}, l_{G}]} C(s) ds - \inf_{s \leq t_G} \int_{[L/v_f, s]} (u(s - L/v_f) - C(s)) ds
\]

As in the case of generalized queuing model, let

\[
h(t) = \int_{0}^{t} (u(s - \tau_0) - C(s)) ds
\]

Leading to

\[
n(t_G) = h(t_G) - \inf_{s \leq t_G} h(s) ds + \int_{[t_G, t_{0}]} u(s) ds
\]
Note that we don’t make any special assumptions on \( t_G \), so Theorem 3 holds for general \( t \). As such, we can characterize the difference between generalized queuing model and the LWR model with a time-dependent boundary. At a single bottleneck, the difference in terms of queue length is reflected in the last residual term in (53). Extension of this result is left to future studies.

4. Numerical experiment

In this section, we examine the performance bounds described in Theorem 2 through numerical experiments. Simulation of a tandem queuing system is straightforward, using an algorithm described Li and Zhang (2015). We include it below for completeness. This algorithm marches over discrete time, and the unit of \( u(t) \) and \( C(t) \) is number of vehicles per time step.

**Algorithm 1 Queue length calculation**

1: **procedure** Queue\((t, u, C)\)  
2: \( s \leftarrow 0, x(s) \leftarrow 0, h(s) \leftarrow 0, g(s) \leftarrow 0 \) \(\triangleright\) Initialization  
3: for \( s = 1 : t \) do  
4: \( h(s) \leftarrow h(s-1) + u(s) - C(s) \)  
5: \( g(s) \leftarrow \min(g(s-1), h(s)) \)  
6: \( x(s) \leftarrow h(s) - g(s) \)  
7: end for  
8: return \( x(t) \) \(\triangleright\) Output  
9: **end procedure**

First, the algorithm is used to simulate single bottleneck dynamics. Upper panel in Figure 3 shows randomly generated model inputs. It models a scenario that bottleneck capacity is increased when a surge in inflow is detected. The lag is 10 time steps. Lower panel of Figure 3 shows the simulated queue length, and its variational structure, meaning that it is nothing else but the deviation between \( h(t) \) and \( \inf_{t \leq t} h(s) \). Note that the well-known negative flow (and thus queue length) issue is also eliminated naturally.

Next, we examine the derived bounds of system performance when flow passes a tandem of bottlenecks. We experiment with three bottlenecks. The capacity of each bottleneck as well as the inflow follow certain random fluctuations, as shown in the upper panel of Figure 4. More precisely, the inflow is defined by

\[
u(t) = \begin{cases} 
35 + 10\epsilon & t \in [41, 50] \\
25 + 10\epsilon & t \in [51, 60] \\
30 + 10\epsilon & \text{otherwise}
\end{cases}
\]

(54)

where \( \epsilon \) follows uniform distribution on \([-0.5, 0.5]\). The perturbation between steps 40 to 60 is to trigger queuing. Bottleneck capacities are i.i.d., equal to \( 30 + 5(\epsilon + 0.5) \). Therefore the system dynamics is stochastic. The lower panel of Figure 4 compares the simulated total queue length in the system and their upper and lower bounds derived in Theorem 2. From the figure, we can tell that these bounds are close approximations to the true value. Moreover, this example also verify our mathematical proof on the tightness of the derived bounds. To more accurately characterize the goodness of approximation of the given bounds, we define two indicators:

\[
\lambda_\ast = \frac{\sum_{j=1}^{J} \max(|\epsilon_U(j)|, |\epsilon_L(j)|)}{\sum_{j=1}^{J} \sum_{n=1}^{N} x_n(j)}, \quad \lambda_\ast = \frac{\sum_{j=1}^{J} \min(|\epsilon_U(j)|, |\epsilon_L(j)|)}{\sum_{j=1}^{J} \sum_{n=1}^{N} x_n(j)}
\]

(55)

where the \( \epsilon_U(j) \) (resp. \( \epsilon_L(j) \)) is the absolute approximation error of upper (resp. lower) bound at time \( j \). These indicators are meant to measure the overall fidelity of the bounds in relative and average sense. We carried simulation experiments to find the value/distribution of \( \lambda_\ast \) and \( \lambda_\ast \), with 1000 simulation runs in each scenario. When three bottlenecks are considered, \( \lambda_\ast \) follows \((0.5374, 0.1548)\) (corresponding to mean and standard deviation), and \( \lambda_\ast \) follow \((0.1658, 0.0733)\); when six bottlenecks are considered, \( \lambda_\ast \) follows \((0.8180, 0.2285)\), and \( \lambda_\ast \) follows \((0.2475, 0.0740)\).
This result is intuitive, which can be interpreted as decreasing predictability when the scale of a stochastic system becomes larger and observed data doesn’t change.

5. Discussion

Modeling queuing dynamics has been a central focus in transportation and other service systems. Traditionally, queuing dynamics are modeled with differential or difference equations, which are ‘local’ in nature, meaning that they capture the instantaneous dynamics, but in general do not directly depict behaviors in the long run, at a large scale, and...
with sufficient analytical insight. This is one of the gaps that the generalized queuing model aims to fill. Formulated in the integral form and possessing a variational solution, this model has attractive analytical and computational properties, and demonstrated here and in other literature (Han et al., 2013; Jin, 2014; Li and Zhang, 2015). Some otherwise puzzling issues, e.g. discontinuity in continuous-time queuing dynamics, have been well addressed in this new modeling framework.

In traffic flow context, there are several analytical tools/models allowing explicit queuing analysis. Ordered with increasing level of generalization, there are: LWR model with piecewise linear $Q_e(\cdot)$ and piecewise constant data (polygon approximation, Dafermos (1972)), LWR model with concave $Q_e(\cdot)$ and constant initial data (Riemann problem, LeVeque (1992)), LWR model with general data with triangle $Q_e(\cdot)$ (simplified kinematic wave theory, Newell

Fig. 4. Upper and lower bounds of total queue length associated with tandem bottlenecks ($n = 3$)
LWR model with general data with concave $Q_\varepsilon(\cdot)$ (the variational formulation, Daganzo (2005)). A tool worth mentioning is the network calculus. Interestingly, though having distinct rationales that are based on two different core concepts (i.e. service curve versus demand surplus), the variational theory and network calculus (max-plus algebra) share some important similarities. For instance, they both do not require the smoothness of inputs, consider cumulative (versus instantaneous) behaviors, and utilize certain variational/optimal properties. This leads to a curious question of whether these two approaches can coalesce to improve the current understanding of queuing dynamics.

It is anticipated that when realistic traffic dynamics on networks are considered, two major challenges exist. First, it is essential to consider routing at network level and ensure properties such as FIFO at link level. Spillover is another complication. In this paper, extending the observation of Nie and Zhang (2002), we demonstrate a way to approximate the kinematic wave model using the point queue model at isolated bottlenecks with time-dependent capacity. To fully address the spillover, it is necessary to model inter-dependency between bottlenecks and queues, which will require more complicated analysis.

6. Conclusions

In this paper, we focus on analytically bounding the performance of a queuing system consisting of tandem bottlenecks. We achieve this goal through establishing a linkage between system-level queuing characters and individual bottleneck dynamics. More specifically, we exploit variational properties of the generalized queuing model, which allows effective concatenation of queuing relations at different locations. Both the mathematical proof and numerical experiment indicate that bounds are tight and provide reasonably accurate estimate of system performance in terms of total queue length. To shed light on realistic traffic systems, we also discuss the relation between the LWR model and the generalized queuing model.

There are several implications. First, this paper demonstrate a novel application of the variational theory pertaining to the point queue process. As we mentioned above, analytical studies of queuing system attributes and dynamics are still limited. This paper points to a new path to address the relevant problems. In particular, using generalized queuing models and exploiting their variational properties appears to be a good starting point to characterize more complex traffic dynamics. Second, from a practical point of view, the upper and lower bounds on queuing system performance allow us to infer system state with limited amount of measurement. This provides substantial benefits for traffic operations, where sensor malfunction and missing data pose significant challenges. Numerical experiment also shows how the derived bounds can be applied in a system with uncertainties. Last but not least, this paper does not rely on the usual Hamilton-Jacobi approach to establish the variational results. Such techniques may be generalized.

In our future work, we are interested to make the following extensions: First, combining network calculus with variational theory, if possible. Although their connection is still not yet clear, it seems that there is a connection between the two and this connection could depend on our understanding of both queuing and traffic systems. Second, addressing the spillover effect in queued networks, which requires a more careful characterization of bottleneck and kinematic wave model. Last but not least, considering how traffic control can be integrated within the present model formulation.

References


Zhang, H., Nie, Y., Qian, Z., 2013. Modelling network flow with and without link interactions: the cases of point queue, spatial queue and cell transmission model. Transportmetrica B: Transport Dynamics 1, 33–51.