Network traffic flow optimization under QoS constraints

Alfréd Csisóka, Themistoklis Charalambousb,c, Hamed Farhadid, Balázs Kulcsárb, Henk Wymeerscb

aInstitute for Computer Science and Control, Hungarian Academy of Sciences, Budapest, Hungary. E-mail: csikos.alfred@aztaki.mta.hu.
bDepartment of Signals and Systems, Chalmers University of Technology, Gothenburg, Sweden. E-mails: {farhad, kulcsar, thecha, henkw}@chalmers.se.
cDepartment of Electrical Engineering and Automation, School of Electrical Engineering, Aalto University, Espoo, Finland.
dJohn A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA.

Abstract

In this paper, a model-based perimeter control policy for large-scale urban vehicular networks is proposed. Assuming a homogeneously loaded vehicle network and the existence of a well-posed Network Fundamental Diagram (NFD), we describe a protected network throughout its aggregated dynamics including nonlinear exit flow characteristics. Within this framework of constrained optimal boundary flow gating, two main Quality of Service (QoS) metrics are considered: (a) first, connected to the NFD, the concept of average network travel time and delay as a QoS metric is defined; (b) second, at boundaries, we take into account additional external network queue dynamics governed by uncontrolled inflow demands. External queue capacities in terms of finite-link lengths are used as the second QoS metric. Hence, the corresponding QoS requirement is an upper bound of external queues. While external queues represent vehicles waiting to enter the protected network, internal queue describes the protected network’s aggregated behaviour.

By controlling the number of vehicles joining the internal queue from the external ones, herewith a network traffic flow maximization solution subject to the internal and external dynamics and their QoS constraints is developed. The originally non-convex optimization problem is transformed to a numerically efficiently convex one by relaxing the QoS constraints into time-dependent state boundaries. The control solution can be interpreted...
as a mechanism which transforms the unknown arrival process governing
the number of vehicles entering the network to a regulated process, such
that prescribed QoS requirements on travel time in the network and upper
bound on the external queue are satisfied. Comparative numerical simula-
tion studies on a microscopic traffic simulator are carried out to show the
benefits of the proposed method.

Keywords: Traffic control; traffic flow; perimeter control; network
fundamental diagram; travel time; Quality of Service.

1. Introduction

Urban traffic congestion has become a major issue, since it results in-
among others - delays, pollutant emissions, higher energy expenditure and
accidents (see, e.g., Bigazzi and Figliozi (2012) and references therein).
Intelligent transportation systems via control and estimation of traffic flows
has been of vital importance to support urban traffic management in order
to appropriately use finite road capacity under different traffic conditions.

One efficient urban traffic coordination approach is to adapt traffic lights
at signalized intersections. To address this problem, several methods has
been deployed at different hierarchical levels (ranging from intersection to
network level), e.g., Papageorgiou et al. (2003). Among these methods,
advanced urban traffic control is one of the most important techniques
aiming at describing urban vehicular networks by some traffic models and
then based on these mathematical abstractions to develop (optimal) con-

trol solutions. Towards this end, the concept of Network Fundamental
Diagram, NFD, often called Macroscopic Fundamental Diagram (MFD),
has been adopted as a basis for the derivation of traffic control strategies
(e.g., Leclercq et al. (2014)). The theory was first proposed in Godfrey
(1969) and further developed in Daganzo and Geroliminis (2008) and Hel-
bing (2009) (its application to experimental data is analyzed in Mahmassani
et al. (1987); Geroliminis and Daganzo (2008); Ampountolas and Kouvelas
(2015); Yildirimoglu et al. (2015)). Daganzo (2007) first used the NFD to
synthesize a controller that maximizes the network outflow, thus compris-
ing a starting point for using the NFD theory for controlling traffic flow.
Several works followed the developed control strategies based on NFD to
maximize the capacity of homogeneous traffic networks. In this case, a
single-region model with one NFD represents the dynamics of the network
appropriately. The paper by Hajiahmadi et al. (2013a) formulates the op-
timal control problem as a mixed integer linear optimization problem, with two types of controllers: perimeter controllers and a switching controller of fix-time signal plans. However, the solution to the problem cannot be used in real time. For alleviating this problem, a Proportional-Integral (PI) controller is proposed by Keyvan-Ekbatani et al. (2012) for real-time gating, with an application to the network of Chania, Greece. By modeling the dynamics of the external queues, a perimeter problem is solved via a Nonlinear MPC formulation in Csikós et al. (2015). Recently, in Haddad and Mirkin (2016), time delays in MFD related control problems have been addressed by means of adaptive control.

Alternative approaches have been used to forecast changing conditions in transportation systems. Due to the complexity of such systems, however, short term travel time estimation and prediction have been in the spotlight for a few decades; see, for example, Vlahogianni et al. (2014) and references therein. We hereby categorise the available techniques according to (1) the regression technique applied (methodology), (2) the type of data (urban or rural), and (3) the source of data collection. First, in order to describe estimation and prediction techniques for travel time, we may follow the model regression methodologies applied. In this vein, both parametric or non-parametric regression techniques have been already suggested. Second, part of the research has been inclined towards the case of the freeway, e.g., Li and Rose (2011) and some to urban Rahmani et al. (2015); Zhan et al. (2013); Jenelius and Koutsopoulos (2013) travel time estimation, or short time prediction. Third, different data sources have been utilized for travel time estimation and predictions (ranging from fixed or mobile sensors to data fusion), see references in Vlahogianni et al. (2014).

The primary goal of the above mentioned works was to inform travellers and hence influence route planning. This means an indirect inclusion of the estimated, predicted travel time. Direct co-design of travel time estimation/prediction information with urban traffic control solution gives rise to improved transportation service, e.g., Lin et al. (2012) proposes a link travel time minimization in a predictive way. Ensuring QoS (e.g., enforced travel time) metrics via traffic control policies is a relevant, non trivial research path, especially in case of large scale traffic networks. In Ramezan et al. (2015), route choice models under dynamical constraint have been included to perimeter flow control decisions. In Haddad (2017) optimal control for maximum queue length inclusion on aggregated inter-regional boundary queues is considered.
In this paper, our main contribution is to develop admission control solutions under multiple QoS requirements: i.e., provide an upper bound average network travel time and keep the external queue size below a certain value. While perimeter flow control has recently received a lot of attention from a control theoretic perspective, further QoS requirements for the system, such as average travel delay in the network, have not been considered. In this work, similarly to the classical perimeter control problem, the objective is to optimize network performance through the maximization of network throughput. However, we additionally include QoS requirements, adopting the service indicators of communication networks (see, for example, Klessig and Fettweis (2014); Liu et al. (2014); Le et al. (2012) and references therein) to (a) keep the travel time spent in the network below a certain threshold, and (b) avoid, if possible, the blockage at the entrance of external queues. These QoS requirements are incorporated as constraints into the gating design procedure. The problem emanating from our objective and constraints, is first formulated as a constrained convex optimization problem and it is solved via internal point methods. If the NFD is assumed to have a quadratic form, the constrained convex optimization problem can be transformed into a QP problem and solved efficiently. Next, we also cast and solve the problem as an MPC one. The performance of our approaches is demonstrated via a case study and compared to that of the proportional-integrator (PI) controller. Perimeter control is implemented on a network implemented on Vissim, a microscopic traffic simulator, describing part of the inner city of Stockholm, Sweden.

The rest of the paper is organized as follows. In Section 2 we provide the necessary notation and preliminaries for the development of our approach. In Section 3 we describe the problem to be targeted and motivate its importance. Then, in Section 4 we present our contributions whose benefits are demonstrated in Section 5 via simulations. Finally, in Section 6 we draw conclusions and discuss possible future research directions.

2. Notation and Preliminaries

The system dynamics is modeled through the conservation of vehicles for both the internal and external queues. The first state equation gives the time evolution of the number of vehicles in the protected/controlled network (representing the evolution of “internal queues”) over a sample step.
of duration $T$, that is,

$$N_{k+1} = \left[ N_k + T \left( \sum_{i \in \mathcal{I}} q_{k}^{\text{in},i} - \sum_{j \in \mathcal{O}} q_{k}^{\text{out},j} + \sum_{h \in \mathcal{D}} d_h^k \right) \right]^+, \quad (1)$$

where $[\cdot]^+$ is the maximum between zero and its argument, $N_k$ denotes the number of vehicles in the network at discrete time-step $k$, $q_{k}^{\text{in},i}$ and $q_{k}^{\text{out},j}$ denote the inflow at link $i$ and outflow at link $j$ at time-step $k$ in unit [veh/h], respectively. $\mathcal{I}$ denotes the set of entrance queues and $\mathcal{O}$ denotes the set of exit links. Let $\mathcal{D}$ denote the set of entrance gates that are ungated. Then, $d_h^k$ is the ungated but measured inflow from the entrance set $\mathcal{D}$, which cannot be compensated by the gated entry links.

Let $q_{k}^{\text{in}} \triangleq \sum_{i \in \mathcal{I}} q_{k}^{\text{in},i}$ and $q_{k}^{\text{out}} \triangleq \sum_{j \in \mathcal{O}} q_{k}^{\text{out},j}$, $d_k \triangleq \sum_{h \in \mathcal{D}} d_h^k$ equation (1) can be abstracted to a single internal queue, i.e.,

$$N_{k+1} = \left[ N_k + T d_k + T \left( q_{k}^{\text{in}} - q_{k}^{\text{out}} \right) \right]^+. \quad (2)$$

The network outflow is modelled through the NFD concept, giving overall network flow $Q$ as a concave function of network accumulation $N$. The total regional circulating flow $Q(N)$ is approximated by Edie’s generalized definition of flow, i.e., the weighted average of link flows multiplied with link lengths. If we assume that the average trip length $\bar{Y}$ in the network is constant and the average link length is given by $l$, then the output (throughput) of the network can be expressed as follows Daganzo (2007):

$$q_{k}^{\text{out}} = \frac{l}{\bar{Y}} Q(N_k). \quad (3)$$

Output flow $q_{k}^{\text{out}}$ is the estimated rate at which vehicles complete trips per unit time either because they finish their trip within the network or because they move outside the network. This function describes steady-state behavior of single-region homogeneous networks if the input to output dynamics are not instantaneous and any delays are comparable with the average travel time across the region Kulcsar et al. (2015).

**Assumption 1.** The function $Q(N)$ is continuously differentiable and concave NFD over the eventual interval on $N$ and network flow is uniform.

Network inflow $q_{k}^{\text{in}}$ is considered to be the controlled input of the system
that follows the admission control policy. This flow depends on the external
queue state, network state, QoS requirements, and the network NFD. The
admittance into the network is described through a simple queuing model,
for entrance gate \(i\), by:

\[
L_{k+1}^i = \left[ L_k^i + T \left( \lambda_k^i - q_{k}^{in,i} \right) \right]^{+},
\]

(4)

where \(L_k^i\) is the queue length of the \(i\)th external queue and \(\lambda_k^i\) denotes
the uncontrolled arrival rate at time \(k\). We assume the arrival rate is an
unknown, deterministic and bounded demand sequence. Exploiting that no
negative queues may appear, by summing all external queues \(i \in \mathcal{I}\),

\[
L_{k+1} = \left[ L_k + T \left( \lambda_k - q_{k}^{in} \right) \right]^{+},
\]

(5)

where \(L_k = \sum_{i \in \mathcal{I}} L_k^i\) and \(\lambda_k = \sum_{i \in \mathcal{I}} \lambda_k^i\). Regarding the overall system, \(d_k\) and
\(\lambda_k\) are considered as measured disturbance.

3. Problem statement

Similar to the classic perimeter control problem, the objective is to op-
timize network performance through the maximization of network through-
put. Moreover, the network performance is characterized by the QoS re-
quirements set. These QoS requirements are usually specified for stochastic
variables, e.g., the expected value of time delay, blockage probability of
external queues. In this work, however, QoS indicators are handled as de-
terministic values. By specifying upper/lower bounds for the indicators,
hard constraints can be given for the system. For the traffic networks,
two QoS requirements are considered: (i) the average time delay in net-
work should be less than a given threshold and (ii) the blockage of external
queues should be avoided.

3.1. Average time delay in network

Suppose the network involves \(M\) distinct links. The average time delay
is modeled by the following formula:

\[
\Delta(N_k) = \frac{l}{v(N_k)} - \frac{l}{v_{\text{free}}},
\]

(6)
where \( l \) denotes the average link length of the network \((l=M^{-1}\sum_{i=1}^{M} l_i)\) for links \( i=1, \ldots, M \), while \( v(N_k) \) and \( v_{\text{free}} \) denote the actual and free link travel speed of the network, respectively.

According to Saberi et al. (2014), the generalized network-wide traffic flow variables, based on the extended Edie’s definitions, can be expressed as follows:

\[
\text{TTD}(N) = Q(N) \cdot T \sum_{i=1}^{M} l_i, \quad (7a)
\]

\[
\text{TTS}(N) = N \cdot T \sum_{i=1}^{M} l_i, \quad (7b)
\]

\[
v(N) = \frac{\text{TTD}(N)}{\text{TTS}(N)}, \quad (7c)
\]

where \( \text{TTD}(N) \) and \( \text{TTS}(N) \) denote the Total Travel Distance and Total Time Spent in the network, respectively and average network speed is expressed as the quotient of these two. Substituting eqs. (7a) and (7b) into (7c), the average network speed can be expressed as follows:

\[
v(N_k) = \frac{Q(N_k)}{N_k}. \quad (8)
\]

Note that \( Q(N_k) \) is chosen such that \( v(N_k) \) is an invertible function. In fact, it is intuitive that as the number of vehicles in the network \( N_k \) increases, the average speed of the network is expected to decrease. In virtue of Assumption 1, invertibility of \( v(N_k) \) is therefore a direct consequence. The free travel speed can be approximated by the following formula:

\[
v_{\text{free}} = \lim_{N_k \to 0^+} \frac{Q(N_k)_{(a)}}{N_k} \equiv \lim_{N_k \to 0^+} \frac{\partial Q(N_k)}{\partial N_k}, \quad (9)
\]

where \((a)\) is due to L’Hôpital’s rule.

Let \( \tau_{\text{free}} \) denote the nominal travel time in the network when a vehicle travels with \( v_{\text{free}} \) and it is equal to \( l/v_{\text{free}} \). We require that the average time delay in the network is smaller than a threshold value, herein denoted by \( \Delta_{\text{tr}}, \) i.e., \( \Delta(N_k+1) \leq \Delta_{\text{tr}}. \)
3.2. Blockage of external queues

A deterministic approach is followed in which the aim is to avoid queue blockage, i.e., \( L_k^i \leq L_{\text{cap},i} \) needs to be satisfied for all \( k \) and \( i \), where \( L_{\text{cap},i} \) denotes the capacity (maximally allowed queue length) of the \( i \)th external queue. This indicator is motivated by the need to avoid gridlocks in the external network through blocking the waiting queues.

**Remark 1.** Arrival rates \( \lambda_i, i \in \mathcal{I} \) are supposed to appear with a similar rate at each entrance link; therefore queues of similar length are built. In the control problem, the sum of all capacities of all external queues are used as a constraint.

**Remark 2.** Note that there may occur such high \( \lambda_i \) rates for which it is not possible to guarantee that both QoS requirements are fulfilled.

4. Main results

In this section, the problem is first cast as an optimization problem. Then, after algebraic manipulations, we restate our constraints (QoS requirements) as upper and lower bounds of the internal queue length.

4.1. QoS conditions

The control aim is to maximize the network outflow (3) such that the specified QoS conditions are satisfied. The outlined QoS conditions can be formalized as follows:

- For the time delay, \( \Delta(N_{k+1}) \leq \Delta_{\text{tr}} \) is given. This condition is used to guarantee a QoS on the travel time vehicles spend in the region. It gives an upper bound for the internal accumulation \( N \) and thus the inflow to the region.

- External queue blockage is avoided if \( L_{k+1} \leq L_{\text{cap}} \). In case of high arrival rates, prescription of this value leads to a lower bound for the internal queue, and indirectly for the inflow to the region.

Additionally, a constraint can be formalized for the admissible flow as follows:

\[
0 \leq q_{k}^{\text{in}} \leq \min(\lambda_k + L_k/T, g_{\max}), \tag{10}
\]

where \( g_{\max} \) denotes the maximal green time of the entering links and it is equal to

\[
g_{\max} = \sum_{i \in \mathcal{I}} g_{\max,i}.
\]
where $g_{\text{max},i}$ denotes the maximal green time of input link $i$. The saturation flow of input links is assumed to be constant (for simplicity of exposition), and it is denoted by $s$. This constraint is not restricting the operation of the network and it basically states that the inflow cannot be less than zero or more than the amount of external queue which can be injected into the network. In the followings, the upper bound for vehicle inflow is denoted by $q^{\text{in,ub}}_k = \min(\lambda_k + L_k/T, g_{\text{max}}s)$.

4.2. Relaxation of the optimal control problem

The general optimization problem for an arbitrary horizon, say of size $m$, can be cast as follows:

\[
\begin{align*}
\max_{q^{\text{in}}_1, \ldots, q^{\text{in}}_m} & \quad \sum_{\ell=1}^{m} Q(N_{k+\ell}) \\
\text{subject to:} & \quad \Delta(N_{k+\ell}) \leq \Delta_{\text{tr}}, \forall \ell \in 1, \ldots, m \\
& \quad \lambda_{k+\ell} = \lambda_k, \forall \ell \in 1, \ldots, m \\
& \quad d_{k+\ell} = d_k, \forall \ell \in 1, \ldots, m \\
& \quad L_{k+\ell} \leq L_{\text{cap}}, \forall \ell \in 1, \ldots, m \\
& \quad 0 \leq q^{\text{in}}_{k+\ell} \leq q^{\text{in,ub}} \forall \ell \in 1, \ldots, m \\
& \quad N_{k+1} = \left[N_k + Td_k + T \left(q^{\text{in}}_k - q^{\text{out}}(N_k)\right)\right]^+ \\
& \quad L_{k+1} = \left[L_k + T \left(\lambda_k - q^{\text{in}}_k\right)\right].
\end{align*}
\]

The multiple step receding horizon control of the above problem leads to a nonlinear optimization problem which is not convex. In the followings, the problem is reformalized as a single step control problem. The special connection between the external and internal dynamics gives basis for relaxing the problem to a convex optimization problem in which the only decision variable is the internal accumulation $N_k$.

The optimization problem can then be formulated as a one step ahead
rolling one as
\[
\begin{align*}
\max_{N_{k+1}} & \quad Q(N_{k+1}) \tag{12a} \\
\text{subject to:} & \quad \Delta(N_{k+1}) \leq \Delta_{tr} \tag{12b} \\
& \quad L_{k+1} \leq L_{\text{cap}} \tag{12c} \\
& \quad 0 \leq q^\text{in}_k \leq q^\text{in,ub}_k \tag{12d} \\
& \quad N_{k+1} = [N_k + Td_k + T(q^\text{in}_k - q^\text{out}(N_k))]^+ \tag{12e} \\
& \quad L_{k+1} = [L_k + T(\lambda_k - q^\text{in}_k)]^+ . \tag{12f}
\end{align*}
\]

We hereby suggest the following optimal delay-aware traffic control policy.

**Proposition 1.** Given a single-step control horizon with constraints (12b)-(12f) on state variables $N_{k+1}$ and $L_{k+1}$ and $q^\text{in}_k$. Optimization problem (12) can be relaxed to a convex optimization problem:
\[
\begin{align*}
\max_{N_{k+1}} & \quad Q(N_{k+1}) \tag{13a} \\
\text{subject to:} & \quad N^\text{lb}_{k+1} \leq N_{k+1} \leq N^\text{ub}_{k+1} \tag{13b}
\end{align*}
\]
from which once the optimization problem (13a) is solved, the optimal control input $q^\text{in}_k$ can be calculated by (2).

**Proof 1.** The upper and lower bounds are obtained as follows. By substituting the speed function $\Delta(N_k)$ from (6) into (12b), a constant lower bound can be derived for the speed, i.e.,
\[
v_{\text{lb, delay}} = \frac{l}{\Delta_{tr} + \tau_{\text{free}}} . \tag{14}
\]
The constant upper bound for the internal queue is obtained by inverting the speed function:
\[
N_{\text{ub, delay}} = v^{-1} \left( \frac{l}{\Delta_{tr} + \tau_{\text{free}}} \right) . \tag{15}
\]
Substituting the upper bound for controlled inflow $q^\text{in}_k$ from (12d) into the equality constraint (12e) a non-constant upper bound emerges and it is given by
\[
N^\text{ub, que}_{k+1} = Tq^\text{in,ub}_k + N_k + Td_k - Tq^\text{out}(N_k) . \tag{16}
\]
As a result, the applied upper bound for the decision variable is given as the minimum of the upper bounds found in (15) and (16), i.e.,

$$N_{k+1}^{ub} = \min(N_{k+1}^{ub,que}, N_{k+1}^{ub,\text{delay}}).$$  

(17)

Lower bound for \(N\) can be obtained by substituting (12e) and (12f) to (12c), i.e.,

$$N_{k+1}^{lb,\text{block}} = N_{k} + Td_k - Tq_k^{\text{out}}(N_{k}) + L_k + T\lambda_k - L_{\text{cap}}.$$  

(18)

Note, that \(N_{k+1}^{lb,\text{block}}\) may take negative values. Hence, the applied lower bound is given as:

$$N_{k+1}^{lb} = \max(0, N_{k}^{lb,\text{block}}).$$  

(19)

**Remark 3.** Due to the min and max functions in the constraint description, we have nonlinear constraints that are usually simplified by a mixed integer formulation. In our approach, time-varying constraints are applied, and assuming a polynomial NFD with a global maximum (e.g., a quadratic function), the maximization of discharge flow leads to a convex optimization problem with new boundary constraints to be solved in each step. As a result, the mixed integer formulation is no longer needed.

Regarding the overall network that involves the external and internal queues, the QoS requirements define a modified capacity of the system through the time varying interval of bounds. As noted in Remark 2, for a very large arrival rate \(\lambda_k\) it is not possible to guarantee that both QoS requirements are fulfilled. This can be seen from (19), where as \(\lambda_k\) increases the lower bound becomes higher, and hence for large \(\lambda_k\) our lower bound may become higher than the upper bound. One of the main advantages of our method, is that it is able to detect when this situation occurs. In such situations, we need to prioritize between the QoS conditions. In our scheme, priority is given to the vehicles in the protected network, i.e., violation of the upper bound, which corresponds to guaranteeing the average time delay in the network, is not permitted. Hence, when the lower bound becomes equal to or even exceeds the upper bound, at that time step the solution of the problem \(N_{k+1}\) is the upper bound itself, and no optimization is required to be solved.

**Proposition 2.** The maximum arrival rate \(\lambda_k\) that can be handled by the network is found by restricting the lower bound of the vehicles in the network
to be smaller than or equal to the upper bound, i.e., \( N^\text{ub}_{k+1} \leq N^\text{ub}_k \). Thus,

\[
\lambda_k \leq \lambda_k^\text{max} \triangleq \max \left( 0, N^\text{ub}_{k+1} + L_{\text{cap}} - N_k - Td_k + Tq^\text{out}_k(N_k) - L_k \right).
\]

For any value above \( \lambda_k^\text{max} \), by choosing \( N_{k+1} \) to be the solution to the optimization, we relax the constraint of having \( L_{k+1} \leq L_{\text{cap}} \) for the external queues in order to keep the network flow at its maximum and avoid compromising the travel delay in the network.

\[ \blacksquare \]

**Proof 2.** The proof directly follows from Proposition 1.

**Remark 4.** In Proposition 2, \( N^\text{ub}_{k+1} \) compresses information on maximum green time and eventual changes in saturation flow.

5. Simulation analysis

In the sequel, the previously proposed macroscopic admission control policy is tested over an emulated traffic network using a microscopic traffic simulator. The case study aims at comparing the following perimeter gating approaches, (i) a simple PI gating policy (see Appendix A.2), (ii) an MPC controller as described in Appendix A.3 and (iii) the proposed relaxed optimal scheme together with the (iv) fixed time strategy (uncontrolled case).

5.1. Simulation environment

For the simulations, the microscopic traffic simulator, Vissim (Fellendorf (1994)), is utilized. Through the COM interface (Tettamanti and Varga (2012)), Vissim is connected to MATLAB, which is used for the online optimization of perimeter signal control. In each cycle, traffic measurements of the states are updated in MATLAB and new control signals are returned to the traffic simulator.

The signal and measurement update cycle of the network are equally 60s. Lengths of external queues are obtained by link measurements. Network accumulation and network average speed are calculated by aggregating individual link data. Network outflow is obtained as the sum of exit flows \( (q^\text{out}_k = \sum_{j \in \mathcal{O}} q^\text{out}_{k,j}) \). The control input is computed as the overall network inflow, \( q^\text{in}_k \), which is then divided to \( q^\text{in}_{k,i} \) entrance flows by following a rule detailed in Appendix A.4.
5.2. Network model

The test network models the city center of Stockholm, Sweden (see Fig. 1).

Figure 1: Layout of the Stockholm network, with entrance links \{I_1, ..., I_{14}\} and exit links \{O_1, ..., O_{15}\}

Traffic enters the network through 14 entrance links, all having the same link capacity \(L_{\text{cap},i} = 50\ \text{veh},\ \forall i \in I\). Network output is served by 15 exit links. In the simulations, fixed routing schemes are used among the origin-destination pairs. Inside the network, a fixed-time signal control is run at all the 24 intersections. Measurements are taken along 78 separate links, measuring average speed and the number of vehicles. The lengths of the longest and the shortest links are approximately 1.98 km and 0.33 km.

For the approximation of the NFD, a quadratic form \(Q(N_k) = aN_k^2 + bN_k\) is applied (see Fig. 2). Parameters \(a\) and \(b\), alongside with further model parameters are given in Appendix A.1, Table 2.

Remark 5. Due to the quadratic NFD, the single-step optimal control presented in Section 3 can be formalized as a QP (quadratic problem). Therefore the relaxed controller for the case study is referred under the label ‘QP’.

5.3. Case study

In the case study a three-hour-long rush hour scenario is analyzed featuring all three controllers. Network load and initial conditions are chosen
such that the conflict of the two QoS criteria can be analyzed. The simulation results are plotted in Figs. 3-6. Fig. 3 depicts the arrival rate and the entrance flows of the different control situations. A sinusoid arrival rate is simulated. In case of no control, the network gets congested around 120min

(see Fig. 4). The three controllers (PI, MPC, QP) however manage to avoid congestions in the protected network.
Figure 4: Number of vehicles in network alongside upper and lower bounds of the QP controller. During 128-180 min, state trajectories of the MPC and QP controllers are zoomed.

The PI control shows a fundamentally different behaviour to that of the MPC and QP controllers. It aims at tracking optimal network accumulation $N_{opt}$. This leads to a very good delay performance, however it entails the blockage of the external queues (Fig. 6). The reason for this is that the state bounds (12b) and (12c) cannot directly be applied to the proposed PI controller. However, the bound for the input signal (12d) is satisfied due to the input saturation (21).

The MPC and QP approaches are very similar regarding the input signal. In the states, however, the different operation during the conflicting QoS requirements can be observed. Starting at 130min, the constraint on travel delay (Fig. 5) and the external queues (Fig. 6) are approached due to the high arrival rates. This causes trouble for the MPC controller as it is not capable of satisfying both constraints, and therefore first violating the delay and then also the blockage constraint. Nevertheless, the QP controller is capable of prioritizing constraints, aiming to keep travel delay below $\Delta_{tr}$ and filling up and thus blocking external queues. At certain points, however, the QP controller seems to violate the state constraint on the internal queue and as a result, the QoS of travel delay (see the zoomed part of Fig.4 and
Fig. 5). As shown in the zoomed plot, in this case, $N_{1h}$ becomes higher than $N_{ulh}$, and the former value needs to be followed as equality constraint. This is not completely satisfied, the QP controller tracks this value with some fluctuation. This is a result of the uncertainty in NFD modeling, as the outflow is not a deterministic function of the internal queue length.

![Graph showing average travel delay per link](image)

Figure 5: Average travel delay per link (calculated from network average speed measurements). The allowed threshold value (see Appendix A.1) for time delay requires the traffic to circulate at least at 50% of the free speed.

Regarding network outflow (Fig. 7), best performance is obtained by the PI controller, however, at the expense of blocking the external queues. There is no significant difference between the MPC and the QP controllers.

Apart from the handling of QoS constraints, the QP controller shows an appealing performance in computational time (see Table 1), due to the relaxation of the problem and the single-step control horizon.

<table>
<thead>
<tr>
<th>Method</th>
<th>PI</th>
<th>MPC</th>
<th>QP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp. time [s]</td>
<td>$6 \times 10^{-4}$</td>
<td>0.241</td>
<td>0.009</td>
</tr>
</tbody>
</table>

Table 1: Computation time of a sample step. (Simulations are run on a PC with Intel i5 3.0GHz CPU and 8GB RAM.)
6. Conclusions and future directions

6.1. Conclusions

In this paper we proposed an admission control mechanism that maximizes network outflow while specified QoS requirements are satisfied. These QoS requirement were incorporated as constraints into the system.
First, a predictive, constrained optimization problem was formulated. Next, the problem was reformulated as a single-step convex optimization problem, and an algorithm was developed ensuring throughput maximization subjected to network travel time constraint guarantees. The performance of our approach was demonstrated via case studies and compared to that of the PI control and MPC control approaches. The case studies illustrate that the proposed mechanism has improved performance in terms of network throughput, average time delay and external queue length.

6.2. Future Directions

Since the shape of the NFD is affected by different factors, it is important to study the problem under uncertain traffic flow description. Towards this end, Kulcsar et al. (2015) proposed an $L_2$ optimal control design, and Haddad and Shraiber (2014) a robust control one, based on the Linear Parameter-Varying (LPV) model structure. However, none of these approaches incorporated QoS requirements, which is part of our ongoing research.

In the case of heterogeneous networks, a set of homogeneous subregions can be defined, described by individual NFDs Ramezani et al. (2015). In this case, multi-region perimeter control is used. In Aboudolas and Geroliminis (2013) perimeter and boundary control is developed via multivariable Linear Quadratic (LQ) regulators. In Hajiahmadi et al. (2013b) the problem of route guidance is solved for a multi-region network. Furthermore, Geroliminis et al. (2013) and Haddad et al. (2013) propose cooperative subregion controllers in a predictive control framework. Current research focuses on extending this work to consider the admission control problem for multiple regions interacting with each other, where each region has as external queues, and hence, gates (part of) the internal queues of other regions, while at the same time, on contrary to existing work in the literature, certain QoS requirements are guaranteed.
A. Appendix

A.1. Network parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>-0.1706</td>
</tr>
<tr>
<td>$b$</td>
<td>872.0880</td>
</tr>
<tr>
<td>$T$</td>
<td>60 s</td>
</tr>
<tr>
<td>$N_{\text{opt}}$</td>
<td>2500 veh</td>
</tr>
<tr>
<td>$N_{\text{max}}$</td>
<td>5000 veh</td>
</tr>
<tr>
<td>$l/\lambda$</td>
<td>0.0111</td>
</tr>
<tr>
<td>$l$</td>
<td>0.6047 km</td>
</tr>
<tr>
<td>$s$</td>
<td>0.5 veh/s/lane</td>
</tr>
<tr>
<td>$y_{\text{max},i}$</td>
<td>40s $\forall i \in \mathcal{I}$</td>
</tr>
<tr>
<td>$v_{\text{nom}}$</td>
<td>42 km/h</td>
</tr>
<tr>
<td>$\tau_{\text{free}}$</td>
<td>51.8 s</td>
</tr>
<tr>
<td>$\Delta_{\tau}$</td>
<td>51.8 s</td>
</tr>
<tr>
<td>$L_{\text{cap},i}$</td>
<td>50 veh $\forall i \in \mathcal{I}$</td>
</tr>
</tbody>
</table>

Table 2: Model parameters

A.2. PI control

The control rule is similar to the one applied in Keyvan-Ekbatani et al. (2012):

$$q_{k}^{\text{in, PI}} = q_{k-1}^{\text{in}} + K_{I}(N_{k} - N_{k-1}) + K_{P}(N_{\text{opt}} - N_{k}),$$  \hspace{1cm} (20)

where $K_{P}$ and $K_{I}$ are the control design parameters, obtained by manual tuning. The design resulted in the following values: $K_{I}=0.3$, $K_{P}=0.07$. Input saturation is applied for the controller in the following form:

$$q_{k}^{\text{in, sat}} = \min(q_{k}^{\text{in, ub}}, q_{k}^{\text{in, PI}}).$$  \hspace{1cm} (21)

where $q_{k}^{\text{in, ub}}$ is given in eq. (10).

A.3. MPC control

MPC is well suited to this problem, since it is a direct constraint handling method that can be implemented over a finite prediction horizon.
Optimization problem (12) is now adapted to MPC framework. First, an equality constraint is involved for the disturbance: throughout the control horizon, $\lambda_k$ is considered constant. Furthermore, the decision variable of the optimization is the vehicle inflow $q_{k}^{\text{in}}$ instead of the internal queue $N_{k+1}$, and distinctively, bounds are defined for the states and the control input. The cost function is also extended. The first term implies the optimization of discharge flow of the protected network. The demand matching in the second term is given to avoid an unnecessary suppression of inflow. The first two terms thus lead to a balanced control of the internal and external queues. The third term is applied to suppress input oscillations. Hence, the optimization problem for the MPC framework is given by

$$\min_{\{q^m_k \ldots q^m_{k+m}\}} \sum_{\ell=1}^{m} \{ - Q(N_{k+\ell}) + \|q_{k+\ell}^{\text{in}} - \lambda_{k+\ell}\|_2^2 \\
+ \|q_{k+\ell}^{\text{in}} - q_{k+\ell-1}^{\text{in}}\|_2^2 \} \quad (22a)$$

subject to

$$N_{k+1} = N_k + Td_k + T\left[ q_{k}^{\text{in}} - q_{k}^{\text{out}}(N_k) \right]^+ \quad (22b)$$

$$L_{k+1} = L_k + T\left[ \lambda_k - q_{k}^{\text{in}} \right] \quad (22c)$$

$$\lambda_{k+\ell} = \lambda_k, \quad \forall \ell \in 1, \ldots, m \quad (22d)$$

$$d_{k+\ell} = d_k, \quad \forall \ell \in 1, \ldots, m \quad (22e)$$

$$0 \leq L_{k+\ell} \leq L_{\text{cap}} \quad \forall \ell \in 1, \ldots, m \quad (22f)$$

$$0 \leq q_{k+\ell}^{\text{in}} \leq \min(\lambda_{k+\ell} + \frac{L_{k+\ell}}{T}, g_{\text{max}}) \quad \forall \ell \in 1, \ldots, m. \quad (22g)$$

The controller solves a convex optimization problem in a rolling horizon manner Grüne and Pannek (2011). For the case study, a control horizon of $N_c = 5$ applied.

**A.4. Division of inflow to multiple entrances**

Arrival rates $\lambda_i$, $i \in \mathcal{I}$ are supposed to appear with an equal rate at each entrance link, therefore queues of similar length are assumed to be built. In spite of this assumption, different queue lengths may be present due to an uneven load of the network. To maintain an equable load of external links, a simple rule is followed to divide input flows, detailed below.

First, define the capacity reserve of external queue $i$:

$$L_k^{\text{res},i} = \max(0, L_{\text{cap},i} - L_k^i). \quad (23)$$
Weighting factor $\alpha_i$ represents the proportion of capacity reserves:

$$\alpha_k^i = 1 - \frac{L_{res,i}^k}{\sum_{i \in I} L_{res,i}^k}$$  \hspace{1cm} (24)$$

where $\beta_i$ gives the fraction of all queued traffic at entrance $i$:

$$\beta_k^i = \frac{L_k^i}{\sum_{i \in I} L_k^i}$$  \hspace{1cm} (25)$$

Overall weighting of inputs are given by $\mu_k^i$ as a combination of $\alpha_k^i$ and $\beta_k^i$:

$$\mu_k^i = \frac{\alpha_k^i \beta_k^i}{\sum_{i \in I} \alpha_k^i \beta_k^i}$$  \hspace{1cm} (26)$$

Controlled inflow $q_k^{in,i}$ at entrance $i$ is then calculated as

$$q_k^{in,i} = \mu_k^i q_k^{in}$$  \hspace{1cm} (27)$$

As a result of the above rule, zero input is given to the entrances with queues of zero length; also, the highest input is given to the queues which are blocked or close to blocking. The rule is proportional to the degree of blockage, giving higher input priority to the links that have more waiting traffic beyond the blocked capacity.

Acknowledgements

This research is supported by Chalmers’ initiatives in transport research, the Transport Area of Advance at Chalmers University of Technology and SAFER (Vehicle and Traffic Safety Centre) and by the National Research, Development and Innovation Office of Hungary - NKFIH through grant No. 115694. B. Kulsoar gratefully acknowledges the support received from Aege.

References


