Delft University of Technology

Delft Center for Systems and Control

Technical report 16-010

Two-level hierarchical model-based predictive control for large-scale urban traffic networks*

Z. Zhou, B. De Schutter, S. Lin, and Y. Xi

If you want to cite this report, please use the following reference instead:

Z. Zhou, B. De Schutter, S. Lin, and Y. Xi, "Two-level hierarchical model-based predictive control for large-scale urban traffic networks," *IEEE Transactions on Control Systems Technology*, vol. 25, no. 2, pp. 496–508, Mar. 2017. doi:10.1109/TCST.2016.2572169

Delft Center for Systems and Control
Delft University of Technology
Mekelweg 2, 2628 CD Delft
The Netherlands
phone: +31-15-278.24.73 (secretary)
URL: https://www.dcsc.tudelft.nl

* This report can also be downloaded via https://pub.bartdeschutter.org/abs/16_010.html

Two-Level Hierarchical Model-Based Predictive Control for Large-Scale Urban Traffic Networks

Zhao Zhou, Bart De Schutter, Shu Lin and Yugeng Xi

***Abstract*—Network-wide control of large-scale urban traffic networks using a hierarchical framework can be more efficient and flexible than centralized strategies for reducing the traffic congestion in big cities, because it can adequately address some problems that occur in controlling such large systems, e.g. computational complexity, multiple control objectives, weak robustness to uncertainties, and so on. In this paper, we propose a two-level hierarchical control framework for large-scale urban traffic networks. At the upper level, based on decomposing a heterogeneous traffic network into several homogeneous subnetworks, a higher-level optimization problem using the concept of macroscopic fundamental diagram is formulated to deal with the traffic demand balance problem. At the lower level, the controller with a more detailed traffic flow model for each subnetwork determines the optimal signal timing within the given region under the guidance of the upper-level controller through communication. For the application of this architecture in real time, the model-based predictive control approach is utilized so as to obtain the best solutions for both levels. Moreover, in order to decrease the computational complexity, a distributed control scheme within each subnetwork is developed at the lower level. The proposed approach is evaluated by simulation under different scenarios on a hypothetical urban traffic network, and the performance is compared with that of other control strategies.**

***Index Terms*—Hierarchical control, model predictive control, large-scale urban traffic networks, macroscopic fundamental diagram.**

I. Introduction

DUE to the rapid development of society and economy, traffic congestion in large-scale urban traffic networks becomes a growing problem all over the world. Therefore, how to deal with the traffic congestion problem on the basis of the available transportation infrastructures is still a serious challenge for the whole society. From a long-term perspective, network-wide traffic signal control is a promising way to alleviate traffic jams.

Different control strategies [1]–[5] with different traffic flow models have been developed for control of urban traffic networks. However, most of these works use centralized control with detailed modeling to obtain a trade-off between the accuracy of modeling and the computational complexity. Since the scale of urban traffic networks consisting of many links and signalized intersections becomes larger, hierarchical or distributed control is more tractable than centralized control for implementation in practice. Hence, a number of researchers have investigated hierarchical architectures for the control of large-scale urban traffic networks. Gartner et al. [6] presented a three-layer control framework, including the signal timings optimization at the local control layer, the offsets optimization for intersections at the coordination layer, and the cycle times calculation at the synchronization layer. A real-time traffic-adaptive signal control system with a three-level hierarchical structure was proposed by Mirchandani and Head [7]. It addressed different problems from the network level to the local level, i.e. network load control, network flow control, and intersection control. De Oliveira and Camponogara [8] proposed a distributed multi-agent framework for control of urban traffic networks by decomposing a centralized control problem into several small coupled sub-problems. Baskar et al. [9] developed a hierarchical control framework inspired by Intelligent Vehicles Highway Systems (IVHS) containing several control levels starting from vehicle controllers at the bottom to supraregional controllers at the top. Moreover, several urban traffic control and management systems have been presented in [10]–[12].

More recently, the concept of macroscopic fundamental diagram (MFD) has been adopted to obtain an efficient and elegant way for control of large-scale urban traffic networks from an aggregated point of view. Its existence was observed and verified by Geroliminis and Daganzo [13] based on experimental data. An MFD links the number of vehicles (or densities) and the space-mean traffic flow in the network. Some theoretical analysis [14], [15] also illustrated that if the variance of link densities is small, i.e., the network is sufficiently homogeneous, the MFD is well-defined, i.e., there is a low scatter of flows for the same density, as shown in Fig. 1. The critical number of vehicles, N_{critical}, corresponding to the maximum space-mean traffic flow, divides the curve of MFD into two parts. The left part of this point corresponds to the uncongested state, in which the average traffic flow is in free flow. The right part corresponds to the congested state, in which the network becomes heavily congested with an increasing number of vehicles. Therefore, if an urban traffic network is considered as a whole, the MFD can describe the characteristics of the network. On the one hand, these findings make it easier to model the dynamics of traffic flow at the network level; on the other hand, researchers can design real-time control strategies based on the MFD to mitigate congestion and to improve mobility in large-scale urban traffic networks. Geroliminis et al. [16] proposed optimal perimeter control for a two-region urban city based on the MFD by regulating the exchanged traffic flows on

Z. Zhou and Y. Xi are with the Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, 200240, China.

B. De Schutter is with the Delft Center for Systems and Control, Delft University of Technology, 2628 CD Delft, The Netherlands.

S. Lin is with the School of Computer and Control Engineering, University of Chinese Academy of Science, Beijing, 100049, China.

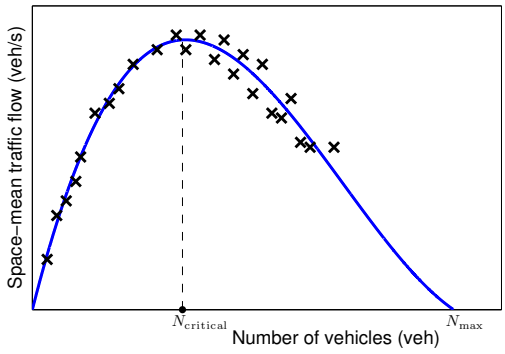

Fig. 1. Well-defined MFD

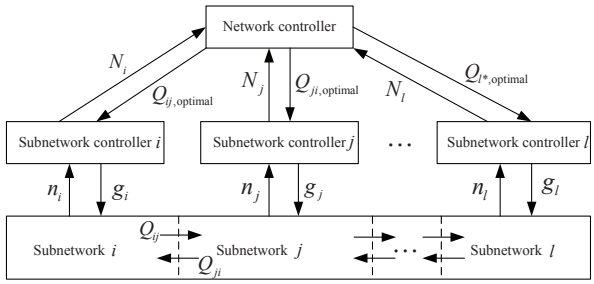

Fig. 2. Hierarchy of two-level MPC

the perimeter borders between the two regions. Boundary control for multiple regions in heterogeneous urban traffic networks has also been investigated in [17]. Moreover, a mixed control strategy integrating perimeter control for urban roads and ramp metering for freeways has been developed in [18], and a hybrid control approach incorporating perimeter controllers and switching signal timing plan controllers has been introduced for urban traffic networks in [19]. Lin et al. [20] developed a high-level controller to regulate the input traffic flows based on the work in [21]. Additional urban traffic control strategies based on the MFD have also been proposed in [22], [23].

In this paper, we propose a two-level hierarchical control framework for large-scale urban traffic networks where at different levels of the hierarchy different models and objectives are taken into account to solve the traffic congestion problem. At each level, model predictive control (MPC) [24] is utilized to solve the optimal control problem, which is a model-based control strategy in which an optimal control sequence is determined by implementing numerical optimization over a given horizon based on a prediction model. Finally, by using a microscopic simulation tool for a large-scale urban traffic network, we show the beneficial properties of the proposed approach compared with other control approaches.

To summarize, this paper contributes to the state-of-the-art in the following ways. First of all, a two-level hierarchical control framework for large-scale urban traffic networks is proposed that is capable of addressing different problems at different layers, i.e. traffic demand balancing and traffic signal coordination. Second, we take fully into account the aggregated characteristics of the MFDs and the task of the upper-level controller, and therefore, we integrate two control performance indicators into the MPC scheme to make control of a multiregion urban network more efficient. Finally, a distributed multi-agent control scheme is presented to decrease the on-line computational complexity of the corresponding optimization problem and to increase the reliability of the controllers at the lower level, which also allows the problem to be solved in a parallel fashion.

The remainder of this paper is organized as follows. In Section II, the structure of two-level hierarchical MPC for large-scale urban traffic networks is discussed. In Section III, a regional traffic flow model based on the MFD is introduced, and then the optimization problem of the upper-level MPC controller is formulated. Then, a distributed multi-agent MPC approach is designed for the lower-level controller in Section IV. Section V presents a simulation-based case study for a typical traffic network. Section VI concludes this paper and outlines future work.

II. Structure of Two-Level Hierarchical MPC

The two-level hierarchical control framework is proposed to improve mobility and to reduce traffic congestion, and further to achieve a better performance for large-scale urban traffic networks. In this framework, we assign the optimization problem to different layers, in which different optimal control problems with specific tasks are solved using the MPC approach. Moreover, the dimension of the optimization problem is also reduced based on a partition of the traffic network, which guarantees that a simpler control problem with a lower dimension can be addressed at a time. Information communication and coordination are considered between the two layers to make the whole system reach a better performance. The control architecture of this approach consists of two layers with one network controller at the upper level and several subnetwork controllers at the lower level, as shown in Fig. 2.

The upper-level controller implements coordination of the lower-level controllers from the network-wide point of view. According to the requirements of the concept of MFD, a heterogeneous large-scale urban traffic network can be decomposed into several homogeneous subnetworks of appropriate scale based on some partition methods [25], [26]. The task of the controller in this layer is to balance the traffic demand among subnetworks and to avoid traffic congestion in each subnetwork. An MFD-based traffic flow model with conservation laws for the space-mean densities and the inflows and outflows of the subnetworks is used as the prediction model, and the optimization problem arising from the MPC approach can then be formulated. This controller performs the optimization by using the current traffic data (the total number of vehicles N_i in subnetwork i, for $i \in \mathscr{M}$, with $\mathscr{M}$ the set of subnetworks) collected from the subnetworks, and then sends the optimal traffic flows to be exchanged among subnetworks ($Q_{ij,\text{optimal}}$, for $j \in \mathscr{N}_i$, with $\mathscr{N}_i$ the set of subnetworks connected to subnetwork i) to the subnetwork controllers as reference targets through communication. It should be noticed that since the traffic model applied in

this layer is comparatively simplistic without considering the dynamic processes inside the subnetworks, the results cannot be applied directly to the traffic signals. Therefore, lower-level controllers are also needed.

At the lower level of the hierarchical structure, each subnetwork controller controls a part of an entire urban traffic network. The aim of the subnetwork controller is to coordinate the intersection signals within the area. Taking into account the instructions given by the upper level, the subnetwork controllers assign the optimal timings g_i for each intersection, so as to regulate traffic flows and mitigate congestion. In this layer, an elaborate traffic model containing more details is used as the prediction model for the MPC optimization problem. In order to reduce the on-line computational complexity and to make the system robust to unexpected disturbances, a distributed multi-agent coordination control approach is proposed. By further dividing each subnetwork into a few subregions, several agents are developed and allocated to the corresponding non-overlapping subsystems. Each agent is capable of making decisions by negotiating with its neighbors for achieving the best performance of the whole system.

III. Upper-level MPC controller

In this section, we focus on the design of the upper-level controller. In order to coordinate the traffic flows among the subnetworks, an aggregate traffic model that can describe the dynamic behavior of the traffic system is needed for the MPC optimization problem. The concept of the MFD provides a tractable way for this modeling. Three characteristics of the MFD have been verified in [13], [27]: (1) there is a unimodal and low-scatter relationship between the network vehicle density and the space-mean flow; (2) the outflow of the traffic network is more or less proportional to the space-mean flow within the network; (3) the shape of the MFD is independent of the traffic demand but is related to the topology of the network and control. Therefore, based on some partition methods [25], [26], a large-scale traffic network can be decomposed into several homogeneous subnetworks with a well-defined MFD. Then we get the following model (see also [16]).

A. Upper-level traffic modeling

Firstly, we can use a simple conservation equation to describe the dynamic evolution of the traffic system for each subnetwork

$$\begin{aligned} N_i(k_u+1) =& N_i(k_u) + T_u\Big(Q_{i,\text{in}}(k_u) - Q_{i,\text{out}}(k_u) \\ &+ \sum_{j\in\mathscr{N}_i}\big(Q_{ji}(k_u) - Q_{ij}(k_u)\big)\Big) \end{aligned} \tag{1}$$

where $N_i(k_u)$ is the number of vehicles in subnetwork i at time step k_u, T_u is the sample time interval, $Q_{i,\text{in}}(k_u)$ is the total inflow from the external origins of subnetwork i, $Q_{i,\text{out}}(k_u)$ is the total outflow to the external exits, $Q_{ji}(k_u)$ is the inflow from subnetwork j, $Q_{ij}(k_u)$ is the outflow exiting to subnetwork j, and $\mathscr{N}_i$ is the set of subnetworks connected to subnetwork i, i.e. the set of neighbors of i.

The MFD defines a static relationship between the number of vehicles and the space-mean flow in the subnetwork:

$$q_i^{\text{w}}(k_u) = \mathscr{G}_i(N_i(k_u)) \tag{2}$$

where $\mathscr{G}_i(N_i(k_u))$ is the function of MFD representing the weighted traffic flow for subnetwork i at $N_i(k_u)$, $q_i^{\text{w}}(k_u)$ is the weighted traffic flow in the subnetwork i:

$$q_i^{\text{w}}(k_u) = \frac{\sum\limits_{r\in\mathscr{R}_i} q_r(k_u) l_r}{\sum\limits_{r\in\mathscr{R}_i} l_r} \tag{3}$$

where $\mathscr{R}_i$ is the set of road segments in subnetwork i, l_r is the length of road segment r, and $q_r(k_u)$ is the traffic flow measured by the corresponding detector in road segment r at time step k_u. Based on traffic data collected from field experiments, an MFD of a network can be extracted. Later on, in the case study of this paper, we will use a polynomial fitting method (the same as the approximation in [16]) to obtain this function.

Finally, for each subnetwork there is a proportional relation between the outflow and the weighted traffic flow based on the second characteristic of the MFD

$$\begin{aligned} D_i(k_u) &= Q_{i,\text{out}}(k_u) + \sum_{j\in\mathscr{N}_i} Q_{ij}(k_u) \\ &= \kappa_i q_i^{\text{w}}(k_u) \end{aligned} \tag{4}$$

where $D_i(k_u)$ is the total outflow of subnetwork i, and κ_i is a coefficient, which can be estimated from real traffic data.

B. Formulation of the upper-level MPC controller

The aim of the upper-level controller is to provide the optimal traffic flow between subnetworks as a reference for the lower-level controllers. Since our goals are to improve the mobility and to maximize the throughput of the subnetworks, the objective function can be defined via two parts: the first is to minimize the number of vehicles in the subnetworks, mitigating the traffic congestion; the second is to keep the number of vehicles in all subnetworks below their critical points, reducing the risk of oversaturation.

Therefore, the total time spent (TTS) is used as the main part of the objective function, which can be described as follows

$$J_{\text{TTS}} = \sum_{i\in\mathscr{M}} \sum_{p=1}^{N_{\text{p}}^{\text{upper}}} N_i(k_u+p)\cdot T_u \tag{5}$$

where $N_{\text{p}}^{\text{upper}}$ is the prediction horizon. In addition, we use another penalty function to meet the second requirement of the control objective

$$J_{\text{Pen}} = \sum_{i\in\mathscr{M}} \sum_{p=1}^{N_{\text{p}}^{\text{upper}}} \left[\max(0, N_i(k_u+p) - N_{i,\text{critical}})\right]^2 \tag{6}$$

where $N_{i,\text{critical}}$ is the critical point for the number of vehicles of the MFD of subnetwork i.

In order to more clearly illustrate this approach, we consider an urban traffic network that has been divided into three subnetworks, i, j, and l, as shown in Fig. 3 (this approach can be extended easily for more subnetworks). By combining the

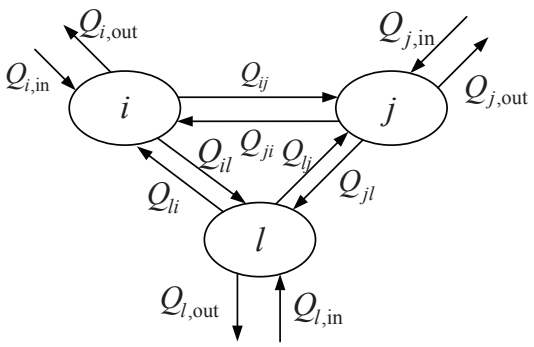

Fig. 3. Three subnetworks

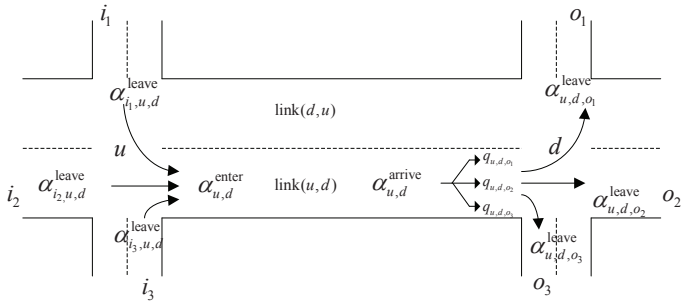

Fig. 4. A link between two adjacent intersections

macroscopic traffic model (1)-(4) and the objective function (5)-(6), we can formulate the upper-level optimization problem as follows

$$
\begin{aligned}
&\min_{\mathbf{Q}(k_u)} \; J_{\text{upper}} = J_{\text{TTS}} + \alpha_{\text{upper}} J_{\text{Pen}} \\
&\text{s.t. } N_i(k_u+p+1) = N_i(k_u+p) + T_u\Big(Q_{i,\text{in}}(k_u+p) \\
&\quad - Q_{i,\text{out}}(k_u+p) + \sum_{j\in\mathscr{N}_i}\big(Q_{ji}(k_u+p) - Q_{ij}(k_u+p)\big)\Big) \\
&\quad D_i(k_u+p) = Q_{i,\text{out}}(k_u+p) + \sum_{j\in\mathscr{N}_i} Q_{ij}(k_u+p) \\
&\quad D_i(k_u+p) = \kappa_i q_i^{\text{w}}(k_u+p) \\
&\quad q_i^{\text{w}}(k_u+p) = \sum_{b=0}^{d} a_b N_i^b(k_u+p) \\
&\quad 0 \le Q_{ij}(k_u+p) \le m_{ij} q_{\text{s},ij} \quad \text{for } j \in \mathscr{N}_i \\
&\quad \text{for } p = 0,\ldots,N_{\text{p}}^{\text{upper}}-1, \text{ for all } i \in \mathscr{M}
\end{aligned}
\tag{7}
$$

where $\mathbf{Q}(k_u) = [Q_{ij}(k_u), Q_{ij}(k_u+1), \ldots, Q_{ij}(k_u+N_p^{\text{upper}}-1)]^{\text{T}}$ for all $i \in \mathscr{M}$ and $j \in \mathscr{N}_i$ is the set of control variables, $\alpha_{\text{upper}} > 0$ is a weighting coefficient, a_b are the coefficients of the MFD polynomial, and d is the polynomial degree. Considering that the traffic system is a real physical system, some constraints have to be imposed on the maximal exchanged traffic flows between subnetworks, i.e., $0 \le Q_{ij}(k_u+p) \le m_{ij} q_{\text{s},ij}$. Here m_{ij} with $j \in \mathscr{N}_i$ denotes the number of links connecting the subnetwork i and its neighbor j, and $q_{\text{s},ij}$ is the average saturation traffic flow for the links between i and j.

IV. Lower-level MPC controller

A. Lower-level traffic modeling

At the lower level, it is necessary to apply a more detailed urban traffic model to regulate the traffic flow within the subnetworks and to track the optimal outflow received from the upper-level controllers. In this section, we use a macroscopic simplified urban traffic model (S model) [5] as the prediction model of subnetwork MPC controllers, for the reason that it can describe the dynamic process of traffic flow in a macroscopic way, including the oversaturated traffic situation.

In the S model, an urban traffic network is constituted of a number of links and intersections. As shown in Fig. 4, a typical urban road (link $(u,d) \in \mathscr{L}$, where $\mathscr{L}$ is the set of links in the whole traffic network) is represented by its upstream intersection u ($u \in \mathscr{E}$, where $\mathscr{E}$ is the set of intersections) and downstream intersection d ($d \in \mathscr{E}$). The sets of the upstream intersections of input links and downstream intersections of output links for link (u,d) are $I_{u,d} \subset \mathscr{E}$ and $O_{u,d} \subset \mathscr{E}$. For the link (u,d) in Fig. 4, we have $I_{u,d} = \{i_1, i_2, i_3\}$ and $O_{u,d} = \{o_1, o_2, o_3\}$. Let $\alpha_{u,d}^{\text{enter}}(k_l)$, $\alpha_{u,d}^{\text{arrive}}(k_l)$, $\alpha_{u,d}^{\text{leave}}(k_l)$ denote the flow rates of vehicles entering link (u,d), the flow rates of vehicles arriving at the tail of the queue in link (u,d) and the flow rates of vehicles leaving link (u,d) at time step k_l, and let $q_{u,d}(k_l)$ be the queue length in link (u,d).

We assume that the cycle time c_{cycle} is equal to the sampling time interval T_l for all intersections. Therefore, the number of vehicles in link (u,d) can be updated by the following conservation equation

$$
n_{u,d}(k_l+1) = n_{u,d}(k_l) + (\alpha_{u,d}^{\text{enter}}(k_l) - \alpha_{u,d}^{\text{leave}}(k_l)) \cdot c_{\text{cycle}} \tag{8}
$$

where the flow rate entering link (u,d) is the sum of the flow rates leaving from its upstream links, i.e.

$$
\alpha_{u,d}^{\text{enter}}(k_l) = \sum_{i\in I_{u,d}} \alpha_{i,u,d}^{\text{leave}}(k_l) \tag{9}
$$

Similarly, the leaving flow rate for link (u,d) is equal to the sum of the flow rates leaving for its downstream links, i.e.

$$
\alpha_{u,d}^{\text{leave}}(k_l) = \sum_{o\in O_{u,d}} \alpha_{u,d,o}^{\text{leave}}(k_l) \tag{10}
$$

The leaving average flow rate over c_{cycle} is determined by

$$
\begin{aligned}
\alpha_{u,d,o}^{\text{leave}}(k_l) = \min(&\beta_{u,d,o}(k_l) \cdot \mu_{u,d} \cdot g_{u,d,o}(k_l)/c_{\text{cycle}}, \\
&q_{u,d,o}(k_l)/c_{\text{cycle}} + \alpha_{u,d,o}^{\text{arriv}}(k_l), \\
&\beta_{u,d,o}(k_l)(C_{d,o} - n_{d,o}(k_l))/c_{\text{cycle}})
\end{aligned}
\tag{11}
$$

where the three terms represent the capacity of the intersection, the number of vehicles waiting and arriving, and the available space in the downstream link, respectively. Moreover, $\beta_{u,d,o}(k_l)$ is the relative fraction of the traffic turning to o at time step k_l, $\mu_{u,d}$ is the saturation flow rate leaving link (u,d), $g_{u,d,o}(k_l)$ is the green time length for the traffic stream towards o in link (u,d), $C_{d,o}$ is the capacity of downstream link (d,o) expressed in number of vehicles, and $n_{d,o}$ is the number of vehicles in link (d,o).

The number of vehicles waiting in the queue turning to o is updated as

$$q_{u,d,o}(k_l+1) = q_{u,d,o}(k_l) + (\alpha^{\text{arriv}}_{u,d,o}(k_l) - \alpha^{\text{leave}}_{u,d,o}(k_l)) \cdot c_{\text{cycle}} \tag{12}$$

After entering the link (u,d), the flow rate of arriving vehicles will reach the tail of waiting queues depending on the turning rates

$$\alpha^{\text{arrive}}_{u,d,o}(k_l) = \beta_{u,d,o} \cdot \alpha^{\text{arrive}}_{u,d}(k_l) \tag{13}$$

For more details about this model, we would like to refer the interested reader to [4], [5].

B. Formulation of the lower-level MPC controllers

The aim of the lower-level controllers is to generate a set of optimal traffic signal timings according to the current traffic conditions. The corresponding algorithm should be embedded in a rolling-horizon framework so that the optimal control problem can be solved on-line before every control cycle.

According to the S model presented in Section IV-A, the dynamic traffic model for each link in subnetwork i ($i \in \mathscr{M}$) can be described as

$$n^i_{u,d}(k_l+1) = f_i(n^i_{u,d}(k_l), g^i_d(k_l), d^i_{u,d}(k_l)) \quad \text{for all } (u,d) \in \mathscr{L}_i \tag{14}$$

where $n^i_{u,d}(k_l)$ is the number of vehicles in link (u,d) of subnetwork i at simulation step k_l, $g^i_d(k_l)$ is the green time of the traffic signals of intersection d, $d^i_{u,d}(k_l)$ is the traffic demand, which can be estimated by using historical data, provided by its neighbors, or received from the upper-level controller, and $\mathscr{L}_i$ is the set of links in subnetwork i.

Since our purpose is to regulate the traffic flows and to guarantee that the outflow of each subnetwork is as close as possible to the optimal exchanged traffic flow between subnetworks, the control objective is to minimize the number of vehicles in the subnetwork and the difference between the real values and the reference set-points for the traffic flows. Therefore, given a prediction horizon $N^{\text{lower}}_{\text{p}}$, the TTS is used as one of the objective functions

$$J_{i,\text{TTS}} = \sum_{(u,d)\in\mathscr{L}_i} \sum_{p=1}^{N^{\text{lower}}_{\text{p}}} n^i_{u,d}(k_l+p) \cdot c_{\text{cycle}} \tag{15}$$

and the second objective function for the lower-level controller can be defined as

$$J_{i,\text{Track}} = \sum_{p=1}^{N^{\text{lower}}_{\text{p}}} \sum_{j\in\mathscr{N}_i} \left(\left(\sum_{e=1}^{m_{ij}} \alpha^{\text{leave}}_{ij,e}(k_l+p) \right) - Q_{ij}(k_l+p) \right)^2 \tag{16}$$

where $\alpha^{\text{leave}}_{ij,e}(k_l+p)$ is the total leaving traffic flow in the links connecting the subnetwork i and its neighbor subnetwork j.

Given the current traffic states at time step k_l measured from all links in the subnetwork i as the initial local states, the future traffic states over a prediction horizon $N^{\text{lower}}_{\text{p}}$ can be predicted as

$$\mathbf{n}^i_{u,d}(k_l) = [n^i_{u,d}(k_l+1|k_l)\ n^i_{u,d}(k_l+2|k_l)\ \cdots\ n^i_{u,d}(k_l+N^{\text{lower}}_{\text{p}}|k_l)]^{\text{T}} \tag{17}$$

Based on the predicted traffic demands

$$\mathbf{d}^i_{u,d}(k_l) = [d^i_{u,d}(k_l|k_l)\ d^i_{u,d}(k_l+1|k_l)\ \cdots\ d^i_{u,d}(k_l+N^{\text{lower}}_{\text{p}}-1|k_l)]^{\text{T}} \tag{18}$$

and the optimal exchanged traffic flows between subnetwork i and its neighbors provided by the upper-level controller, the optimization problem of MPC solved by local lower-level controller can be formulated as follows

$$\begin{aligned} \min_{\mathbf{g}^i(k_l)}\ & J_{i,\text{lower}} = J_{i,\text{TTS}} + \alpha_{\text{lower}} J_{i,\text{Track}} \\ \text{s.t. }\ & n^i_{u,d}(k_l+p+1) = f_i(n^i_{u,d}(k_l+p), g^i_d(k_l+p), \\ & \quad d^i_{u,d}(k_l+p)) \\ & \quad \text{for } p = 0,\ldots,N^{\text{lower}}_{\text{p}}-1, \text{ for all } (u,d)\in\mathscr{L}_i \\ & \Phi(\mathbf{g}^i(k_l)) = 0 \\ & \mathbf{g}^i_{\min} \le \mathbf{g}^i(k_l) \le \mathbf{g}^i_{\max} \end{aligned} \tag{19}$$

where $\alpha_{\text{lower}} > 0$ and $\mathbf{g}^i(k_l)$ contains the optimized control inputs for all the intersections in subnetwork i, i.e.

$$\begin{aligned} \mathbf{g}^i_d(k_l) =& [g^i_d(k_l|k_l)\ g^i_d(k_l+1|k_l)\ \cdots \\ & g^i_d(k_l+N^{\text{lower}}_{\text{p}}-1|k_l)]^{\text{T}} \\ \mathbf{g}^i(k_l) =& [\mathbf{g}^{i\text{T}}_1(k_l)\ \mathbf{g}^{i\text{T}}_2(k_l)\ \cdots\ \mathbf{g}^{i\text{T}}_\phi(k_l)]^{\text{T}} \end{aligned} \tag{20}$$

where ϕ is the number of intersections in subnetwork i, $\Phi(\mathbf{g}^i(k_l)) = 0$ represents the cycle time constraints for all intersections in the subnetwork, and $\mathbf{g}^i_{\min}$ and $\mathbf{g}^i_{\max}$ are the bounds for the green time signals. The cycle time constraints guarantee that the sum of all the green time durations for an intersection equals the given cycle time, i.e. $\sum_{h\in H_d} g_{d,h}(k_l) = c_{\text{cycle}}$, where H_d is the set of number of traffic signal phases at intersection d, and $g_{d,h}(k_l)$ is the green time duration of phase h at intersection d at time step k_l.

C. Distributed multi-agent MPC approach

Due to the high computational complexity and low reliability (e.g., there is a single point of failure in case the single controller breaks down) of centralized MPC, it is necessary to apply a distributed MPC approach for on-line control. In such an approach the overall problem could be decomposed into several subproblems. On the one hand, the computational complexity is significantly reduced because one controller could determine the control actions for its subsystem by solving a low-dimensional optimization problem; on the other hand, this approach could prevent the breaking down of integrated system from the failure of one controller, although the resulting solution will be sub-optimal. However, this approach poses many challenges in practice, such as communication delays, communication errors and so on. It is noted that in particular if agents fail to provide the accurate information to their neighbors, this approach cannot yield a good overall performance. In this subsection, we propose a multi-agent MPC scheme to deal with the control problem for traffic subnetworks.

Assume without loss of generality that an urban subnetwork i can be further decomposed into several subregions (i.e. subsubnetworks) in terms of some network partition methods [25], [26], the capability of processors, or the instructions from

traffic operators. For subregion $w \in \mathscr{M}_i$ ($\mathscr{M}_i$ is the set of subregions in subnetwork i) with its neighbor $v \in \mathscr{N}_w$ ($\mathscr{N}_w$ is the set of the subregions in subnetwork i connected to subregion w, i.e. the set of neighbors of subregion w in subnetwork i), the optimization problem can be expressed as

$$\begin{aligned}
&\min_{\mathbf{g}^{i_w}(k_l)} J_w = J_{i_w,\text{lower}} \\
&\text{s.t. } n^{i_w}_{u,d}(k_l+p+1) = f_{i_w}(n^{i_w}_{u,d}(k_l+p), g^{i_w}_d(k_l+p), \\
&\quad d^{i_w}_{u,d}(k_l+p), z_{vw}(k_l+p)) \\
&\quad \Phi(\mathbf{g}^{i_w}(k_l)) = 0 \\
&\quad \mathbf{g}^{i_w,\min} \le \mathbf{g}^{i_w}(k_l) \le \mathbf{g}^{i_w,\max} \\
&\quad z_{vw}(k_l+p) = y_{vw}(k_l+p) \\
&\quad \text{for } p = 0,\ldots,N^{\text{lower}}_{\text{p}} - 1 \\
&\quad \text{for all } (u,d) \in \mathscr{L}_{i_w}, \text{ for } v \in \mathscr{N}_w
\end{aligned} \tag{21}$$

where $y_{vw}(k_l+p)$ and $z_{vw}(k_l+p)$ are the interaction variables between subregions w and v. More specifically, $y_{vw}(k_l+p)$ is the output flow of subregion v and then into subregion w, and $z_{vw}(k_l+p)$ is the input flow of subregion w from v. Obviously, the interaction traffic flow $z_{vw}(k_l+p)$ must be equal to $y_{vw}(k_l+p)$. Therefore, the interactions between subregions will be guaranteed by the following interaction constraints

$$y_{wv}(k_l+p) = z_{wv}(k_l+p) \tag{22}$$
$$y_{vw}(k_l+p) = z_{vw}(k_l+p) \tag{23}$$

However, since each interaction constraint contains two variables from different agents, it cannot be added into the optimization problem of any of the individual agent directly. Therefore, in order to make sure the satisfaction of interaction constraints among subregions, the coordination methodology of multi-agent MPC is developed.

The combined overall control problem of subnetwork i is formed by the aggregation of the local agents (21) and the interaction constraints (22)-(23). Due to the interaction constraints, this problem is not able to be decomposed into several independent optimization subproblems using only local information. In order to deal with this problem, the dual decomposition method (the augmented Lagrangian method) [28]–[30] is introduced to move the interaction constraints into the objective function in the form of using the Lagrangian multipliers and additional quadratic terms. Therefore, the Lagrangian function of the overall optimization problem can be written as

$$\begin{aligned}
L_i = \sum_{w\in\mathscr{M}_i} \Bigg(J_w + \sum_{v\in\mathscr{N}_w} \sum_{p=0}^{N^{\text{lower}}_{\text{p}}-1} \Big(&\lambda_{vw}(k_l+p)(z_{vw}(k_l+p) \\
&- y_{vw}(k_l+p)) + \frac{c}{2}\|z_{vw}(k_l+p) - y_{vw}(k_l+p)\|_2^2 \Big) \Bigg)
\end{aligned} \tag{24}$$

where $\lambda_{vw}(k_l+p)$ is the Lagrangian multiplier vector corresponding to the interaction constraint $z_{vw}(k_l+p) = y_{vw}(k_l+p)$, and c is a positive constant.

Since the formulation (24) includes the non-separable quadratic terms, we approximate it with the following equation by using the approach proposed in [31]:

$$\begin{aligned}
\widetilde{L}_i &= \sum_{w\in\mathscr{M}_i} (J_w + J_{w,\text{inter}}) \\
&= \sum_{w\in\mathscr{M}_i} \Bigg(J_w + \sum_{v\in\mathscr{N}_w} \sum_{p=0}^{N^{\text{lower}}_{\text{p}}-1} \Bigg(\begin{bmatrix} \lambda^s_{vw}(k_l+p) \\ -\lambda^s_{wv}(k_l+p) \end{bmatrix}^{\text{T}} \begin{bmatrix} z_{vw}(k_l+p) \\ y_{wv}(k_l+p) \end{bmatrix} \\
&\quad + \frac{c}{2} \left\| \begin{bmatrix} z^{s-1}_{wv}(k_l+p) - y_{wv}(k_l+p) \\ y^{s-1}_{vw}(k_l+p) - z_{vw}(k_l+p) \end{bmatrix} \right\|_2^2 \Bigg) \Bigg)
\end{aligned} \tag{25}$$

where $J_{w,\text{inter}}$ is the cost function associated with the interaction variables. At each iteration s, the variables $\lambda^s_{vw}(k_l+p)$ and $\lambda^s_{wv}(k_l+p)$ are the Lagrange multipliers for its interaction constraints $z_{vw}(k_l+p) = y_{vw}(k_l+p)$ and $z_{wv}(k_l+p) = y_{wv}(k_l+p)$ respectively. Moreover, $z^{s-1}_{wv}(k_l+p)$ and $y^{s-1}_{vw}(k_l+p)$ are the previous information of the agents of the last iteration $s-1$.

The approximation of the non-separable quadratic terms is solved by using the so-called auxiliary problem principle [32]. Compared with the popular alternating direction method of multipliers [33], this approach allows agents to address an approximation of the augmented Lagrangian problem in a parallel way. In order to reduce the number of optimization variables of the control problem of each agent, the input traffic flow of subregion w at every iteration s, i.e. $z^s_{vw}(k_l+p)$, is estimated using the previous information from its neighbors, e.g. $z^s_{vw}(k_l+p) = y^{s-1}_{vw}(k_l+p)$. Therefore, the distributed multi-agent MPC approach for urban traffic subnetworks at each control step k_l can be described in a flowchart, as shown in Fig. 5, where $\mathbf{e}^s = z^s_{vw}(k_l+p) - y^s_{vw}(k_l+p)$ is the error between the desired traffic flow input $z^s_{vw}(k_l+p)$ and the real traffic flow supply $y^s_{vw}(k_l+p)$ from the neighboring subregions, and $\varepsilon > 0$ is a threshold value.

V. Simulation-based case studies

To evaluate the effectiveness of the proposed two-level hierarchical MPC method for urban traffic management, we build a hypothetical urban traffic network to assess the performance of the proposed approach and to compare it with other existing control approaches, namely fixed-time control, centralized MPC, and decentralized MPC.

A. Scenarios and set up

The simulated network is shown in Fig. 6. There are 55 nodes including 21 source nodes providing traffic demands and 34 intersections controlled by traffic signals, and 154 two-way links with different lengths (216-366 m). All the links have two lanes. The simulation is carried out by using CORridor SIMulation (CORSIM), C++, and MATLAB. CORSIM is a microscopic traffic simulation tool for implementing traffic operations. The rolling-horizon optimization problem is solved in MATLAB, while C++ provides the interface between CORSIM and MATLAB. Considering that the optimization problems at both levels are non-linear non-convex problems because of the nonlinearity of the models,

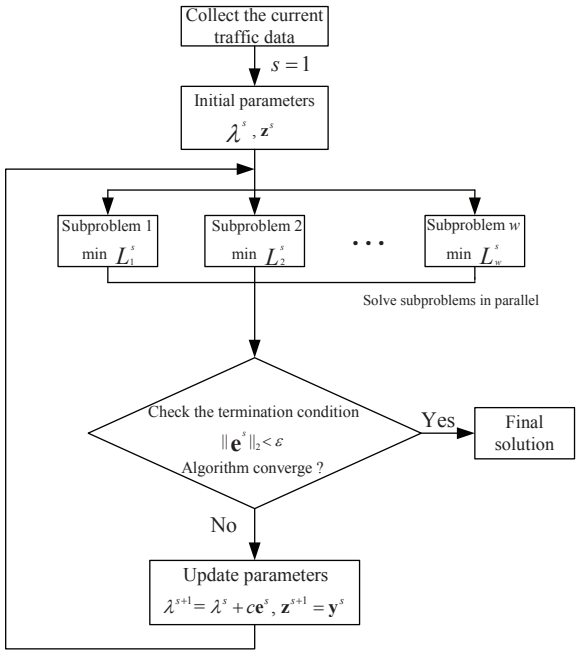

Fig. 5. Flowchart of the distributed multi-agent MPC algorithm

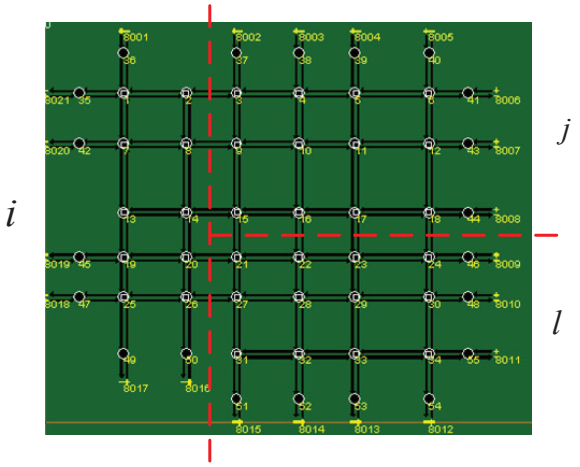

Fig. 6. Urban traffic network used for simulation

the function *fmincon* in the MATLAB optimization toolbox based on Sequential Quadratic Programming (SQP) is utilized to calculate the optimal control inputs. Moreover, in order to avoid the optimization ending up in a local minima, a multi-start technique is used to search for a global optimal solution. Given different random initial points, we run the solver for each one and record the results. The one corresponding to the lowest objective function value is selected as the optimal solution and then applied to the traffic signals. With respect to the selection of the number of initial points, please refer to [34]. Therefore, five initial points are adopted in the multi-start SQP approach for the following case studies.

First of all, the entire urban traffic network should be appropriately divided into several subnetworks. We consider the partition method proposed by Zhou et al. [26] and divide the whole network into three subnetworks, as shown in Fig. 6. In order to implement distributed MPC control at the lower level, each subnetwork is partitioned into two subregions with the same size by taking into account the computational efficiency. Since the shape of the MFD is related to the signal timing plans, we execute 5 predefined fixed-time plans for the signalized intersections in each subnetwork. The selection of these plans is based on the tuning green time ratio for the intersections, where each plan is predefined for the undersaturated (2 plans), saturated and oversaturated traffic situation (2 plans). Therefore, the MFDs under different signal timing plans for each subnetwork can be obtained, as shown in Fig. 7(a), (b) and (c), respectively. In this case study, in order to obtain the relationship in (2), we use a five-order polynomial function of number of vehicles in each subnetwork to derive an averaged nonsymmetric unimodal MFD, e.g., $q^{\mathrm{w}}(k_u) = a \cdot N^5(k_u) + b \cdot N^4(k_u) + c \cdot N^3(k_u) + d \cdot N^2(k_u) + e \cdot N(k_u) + f$, where a, b, c, d, e, and f are estimated parameters with unit $/s$. From Fig. 7(a), (b) and (c), the parameters for all three subnetworks are obtained by using the same approximation method in [16], i.e., $a_i = 8.1374 \times 10^{-17}$, $b_i = -6.0171 \times 10^{-13}$, $c_i = 1.6470 \times 10^{-9}$, $d_i = -2.1778 \times 10^{-6}$, $e_i = 0.0014$, $f_i = -0.0899$, $a_j = 8.6971 \times 10^{-18}$, $b_j = -1.1286 \times 10^{-13}$, $c_j = 5.0639 \times 10^{-10}$, $d_j = -1.0573 \times 10^{-6}$, $e_j = 0.0011$, $f_j = -0.0936$, $a_l = 9.2148 \times 10^{-19}$, $b_l = -4.1038 \times 10^{-14}$, $c_l = 2.8660 \times 10^{-10}$, $d_l = -7.9809 \times 10^{-7}$, $e_l = 8.8556 \times 10^{-4}$, $f_l = -0.0468$, $N_{i,\mathrm{critical}} = 1300$ veh, $N_{j,\mathrm{critical}} = 1700$ veh, $N_{l,\mathrm{critical}} = 1000$ veh. Finally, to estimate the relationship in (4), we calculate the weighted traffic flow q^{w} and the total outflow D for all three subnetworks based on the traffic data from CORSIM. The results are shown in Fig. 7(d), (e) and (f), i.e., $\kappa_i = 0.08$, $\kappa_j = 0.07$, $\kappa_l = 0.075$.

In this paper, we consider two scenarios with different traffic demands (i.e. the network input flow rates) for simulation of the network. The first corresponds to an increasing, high traffic demand to simulate the oversaturated traffic condition. The second is to simulate a peak hour situation with an increasing demand from the beginning of the simulation and then a decreasing demand towards the end. For simplicity, the input traffic flow rates of all the source nodes to the network are assumed to be equal, and the traffic demand variation is illustrated in Table I. The cycle times of the traffic signals c_{cycle} are 60 s for all intersections and the offsets between two adjacent intersections are 0 s during the simulation. These two parameters are constant in our simulations. The sample time interval adopted at the upper and lower level is the same and is equal to the cycle time of signalized intersections, i.e. $T_u = T_l = 60$ s. The total simulation time is 5400 s. The control time interval T_{c} is 180 s, which corresponds to 3 sample time intervals. At each intersection, the turning rate $\beta_{u,d,o}$ for each direction is 33.33%. The saturation flow rate $\mu_{u,d}$ is 2000 veh/h per link. The lower and upper bounds of the green time signals are selected as $g_{\min} = 10$ s and $g_{\max} = 50$ s for all intersections. The prediction horizons are the same for all the MPC controllers, and for both levels $N_{\mathrm{p}}^{\mathrm{upper}} = N_{\mathrm{p}}^{\mathrm{lower}} = 7$ (i.e. 21 min). (The choice for the prediction horizon is based on the analysis in [34]). Moreover, the weighting coefficient $\alpha_{\mathrm{upper}} = 10^{-4}$ in (7) and the weighting coefficient $\alpha_{\mathrm{lower}} =$

10^{-3} in (19) are obtained based on the nominal values of the main objective function and the penalty or tracking term.

TABLE I
NETWORK INFLOW FOR EACH SOURCE NODE

Simulation time (s)	Traffic demand flow (veh/h)	
	Scenario 1	Scenario 2
0-900	2000	800
900-1800	2000	1000
1800-2700	2500	1200
2700-3600	2500	1500
3600-4500	3000	1500
4500-5400	3000	1000

In the following we compare four control methods:

1) Fixed-time control method, which is a signal control plan where the green time split has been predefined for each intersection. Here, we use the best one of the 5 predefined fixed-time plans, i.e., the cycle times is 60 s for all intersections, and the green time durations are 30 s and 30 s for the two phases of all the intersections.
2) A single agent using centralized MPC to control the whole network. In this approach, a large-scale urban traffic network is controlled by a single centralized agent without decomposition into several subnetworks. Moreover, the S model is utilized as the prediction model of the MPC controller. At each control step, the MPC optimization problem is formulated by minimizing the TTS in the network subject to the dynamics of network over the horizon and the input constraints.
3) Decentralized MPC strategy. The whole urban traffic network is decomposed into several subnetworks. Each subnetwork is assigned an MPC controller. However, there is no communication between one controller and its neighbors. In other words, when the decentralized strategy is adopted, each subnetwork controller solves an independent optimization problem without the information (traffic demands) provided by its neighbors and without the coordination by the upper-level controller. Therefore, the optimization problem is formulated by minimizing the objective function TTS in (15) subject to the local dynamics of subnetwork over the horizon and the input constraints in (19).
4) Hierarchical control based on the proposed two-level coordinated MPC approach. This approach involves the upper-level controller described in Section III and the lower-level control strategy described in Section IV.C.

In order to compare the results and to evaluate the performance of each control approach, we consider four estimation criteria. $\mathrm{TTS}_{\mathrm{eval}}$ is the accumulated amount of the total time spent by all the vehicles inside the traffic network since the beginning of the simulation, including both the vehicles running freely on a link and the vehicles slowing down or waiting in queues

$$\mathrm{TTS}_{\mathrm{eval}} = \sum_{k=1}^{K_c} \sum_{(u,d)\in\mathscr{L}} T_c \cdot n_{u,d}(k) \tag{26}$$

where K_c is the number of control time steps in the considered time horizon. Total Delay Time (TDT) [5] is the difference between the total travel time of all vehicles inside the road network since the beginning of the simulation and the total free-flow travel time (i.e., the time needed by the vehicles traveling at the maximum permitted speed), so the TDT is actually the total amount of time that the vehicles are delayed:

$$\mathrm{TDT} = \sum_{k=0}^{K} \sum_{(u,d)\in\mathscr{L}} \left(\frac{l_{u,d}}{v_{u,d}^{\mathrm{average}}(k)} - \frac{l_{u,d}}{v_{u,d}^{\mathrm{free}}} \right) \cdot n_{u,d}(k) \tag{27}$$

where $l_{u,d}$ is the length of link (u,d), $v_{u,d}^{\mathrm{average}}(k)$ is the average speed of all vehicles in link (u,d) at time step k. It is provided directly by CORSIM. The weighted average flow has been defined in (3), which represents the mobility of the traffic network. The number of vehicles in each subnetwork is also considered as a criterion.

B. Simulation results

In this section, simulation results are presented to assess the efficiency for two traffic scenarios of the two-level hierarchical MPC control approach compared with the other three control methods. All data collected from CORSIM is used to evaluate the performance.

The $\mathrm{TTS}_{\mathrm{eval}}$ and TDT results of entire traffic network for all control approaches in the two scenarios are shown in Fig. 8. From Fig. 8(a), we can see that the centralized, the decentralized and the hierarchical control approaches yield a significant decrease in the $\mathrm{TTS}_{\mathrm{eval}}$ compared with fixed-time control. Before the 20th control step, the traffic network is not very congested. The difference between the three control approaches is not obvious. When the traffic situation reaches the oversaturated condition because of the high traffic demand, the centralized control exhibits a better performance than the other two control approaches. The average difference between decentralized control and centralized control is 3.78% and the maximal difference is 8.52%. The average difference between hierarchical control and centralized control is 2.46% and the maximal difference is 4.8%. Although the optimization objective of decentralized control is TTS, the MPC controllers only consider the local information in their own subnetwork without communication with their neighbors. This will not result in the global optimum for the whole network. From Fig. 8(b), we can see that compared to decentralized control, the improvement in TDT by using hierarchical control is more obvious, especially in the oversaturated traffic situation. This means that hierarchical control is capable of improving the mobility in the network. At the upper level, hierarchical control not only considers minimization of the objective function TTS, but also takes the MFDs of the subnetworks into account to prevent the traffic situation from falling into the oversaturated condition. According to the guidance from the upper level, the lower-level MPC controllers are able to coordinate the traffic flows through communication, and then achieve a better performance for the whole network. The results shown in Fig. 8(c) and 8(d) under Scenario 2 also confirm the efficiency of our proposed method and the conclusions. The detailed comparisons are shown in Table II. The results illustrate that hierarchical control can approximate the performance of

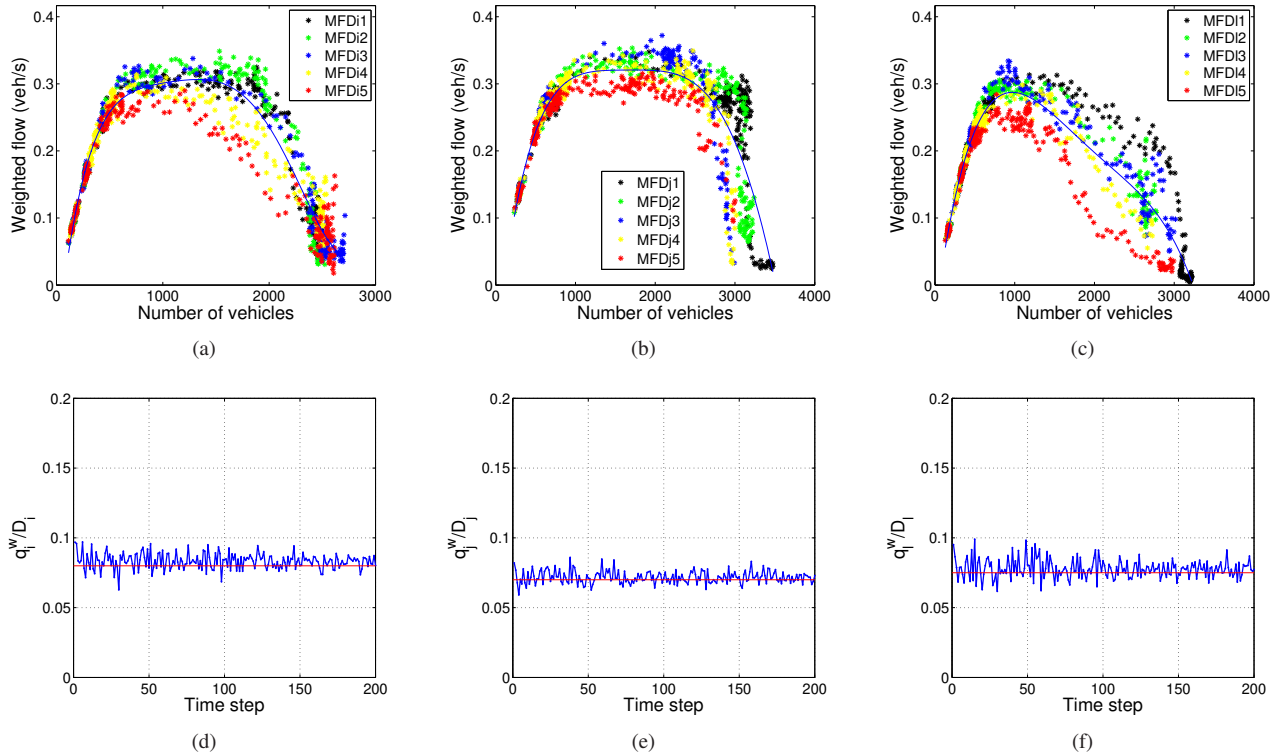

Fig. 7. Characteristics of MFDs for the different subnetworks. (a), (b) and (c) Relationship between the number of vehicles and the weighted average flow. (d), (e) and (f) Relationship between the weighted average flow and the total output.

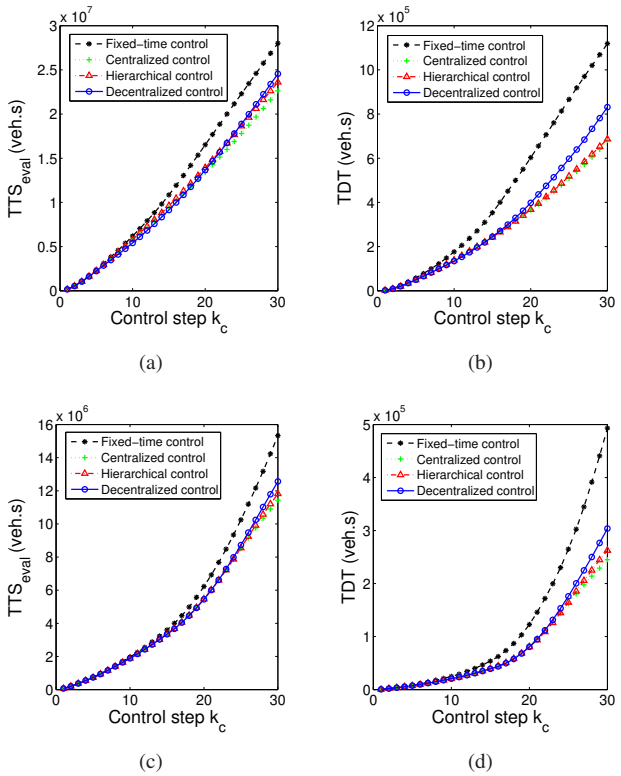

Fig. 8. TTS_{eval} and TDT comparison for all control approaches in the two scenarios. (a) and (b) Scenario 1. (c) and (d) Scenario 2.

centralized control, and is able to reduce the TTS and TDT more than decentralized control.

In order to investigate that hierarchical MPC control is able to balance the traffic demand and to coordinate the traffic flow, we further compare the weighted average flow and the number of vehicles in each subnetwork by different control strategies under the two scenarios. The evolution of the weighted average flow over time corresponding to three subnetworks i, j and l in Scenario 1 is shown in Fig. 9(a), 9(b) and 9(c), respectively. These figures demonstrate that hierarchical MPC control and centralized MPC control are both able to keep a relative higher traffic flow in each subnetwork compared with the other two control schemes in an increasing traffic demand situation. This can be inferred from the evolution of the number of vehicles in each subnetwork shown in Fig. 9(d), 9(e) and 9(f). From Fig. 9(d), we can see that fixed-time control and decentralized control lead the number of vehicles in subnetwork i to the jam state, and that hierarchical control keeps the number of vehicles near the critical point $N_{i,\text{critical}} = 1300$ veh, preventing the traffic situation from achieving the oversaturated situation, and that centralized control performs a little worse in this aspect than hierarchical control. The same result also appears in Fig. 9(f). The average differences between the number of vehicles and the critical point under hierarchical control are less than the differences under centralized control for subnetwork i and l. However, Fig. 9(e) shows that the number of vehicles in subnetwork j under hierarchical control is larger than the number of vehicles under centralized control, i.e. hierarchical control performs worse than centralized control

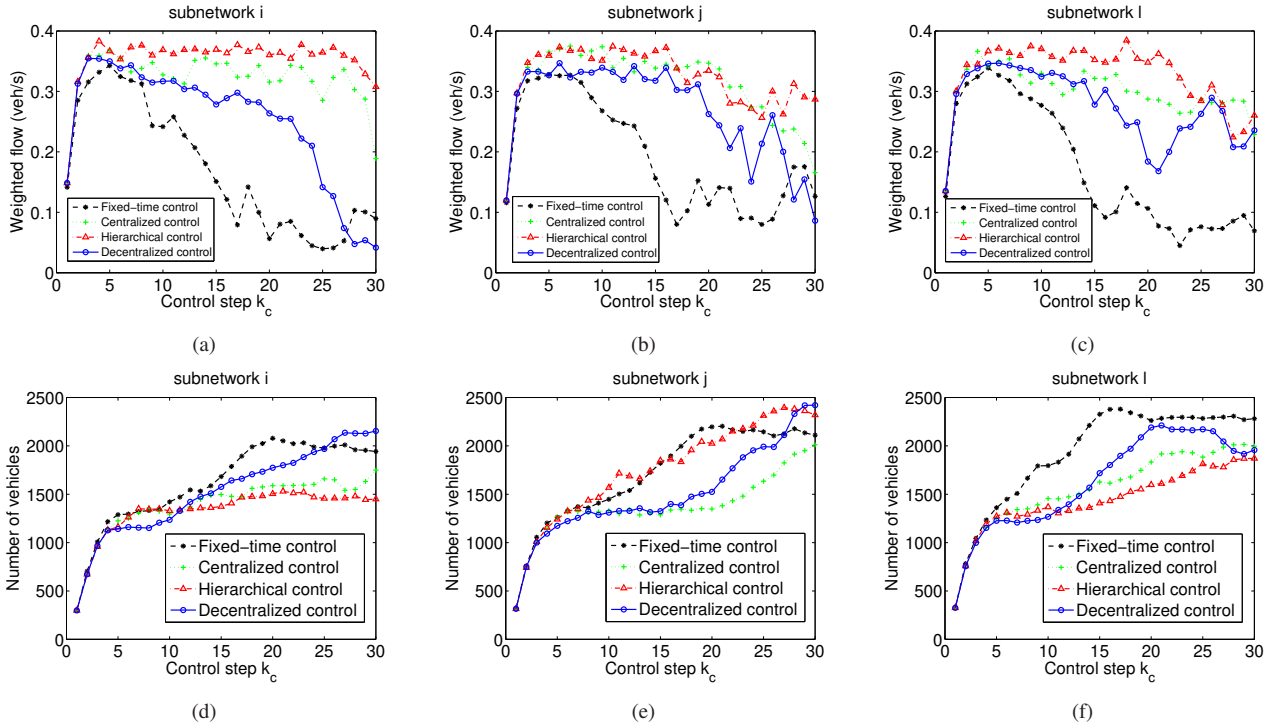

Fig. 9. Comparison for two performance indices of the four control strategies for three subnetworks in Scenario 1. (a), (b) and (c) Weighted average flow. (d), (e) and (f) Number of vehicles.

TABLE II
TTS_{eval} AND TDT FOR ALL CONTROL APPROACHES IN THE TWO SCENARIOS

Control approach	S1				S2			
	TTS_{eval} (veh·s)	Improve (%)	TDT (veh·s)	Improve (%)	TTS_{eval} (veh·s)	Improve (%)	TDT (veh·s)	Improve (%)
Fixed-time	2.80×10^7	-	11.2×10^5	-	1.53×10^7	-	4.93×10^5	-
Centralized	2.26×10^7	19.3	6.81×10^5	39.2	1.14×10^7	25.5	2.45×10^5	50.3
Decentralized	2.46×10^7	12.1	8.31×10^5	25.8	1.26×10^7	17.6	3.04×10^5	38.3
Hierarchical	2.36×10^7	15.7	6.87×10^5	38.7	1.18×10^7	22.9	2.62×10^5	46.9

in subnetwork j. This can be explained by the fact that the MFD of subnetwork j has a wide saturation area as shown in Fig. 7(b), and also because of the upper-level controller balancing the traffic demand among three subnetworks, which makes a compromise for the improvement of performance of the whole network. In the absence of network-wide real-time optimization, fixed-time control leads to an oversaturated traffic situation, and even to a gridlock situation. Without communication and coordination, decentralized control takes the risk of increasing the congestion degree of the whole network. Centralized control can achieve the best performance of the whole system, however, hierarchical control not only can approximate the performance of centralized control, but also is capable of balancing the distribution of number of vehicles of road network in an increasing traffic demand situation.

Furthermore, we also compare the weighted average flow and the number of vehicles in each subnetwork under Scenario 2, i.e., the peak hour simulation. From Fig. 10(a), 10(b) and 10(c), we can see that all three MPC schemes, i.e. centralized control, decentralized control and hierarchical control, yield a high average traffic flow in three subnetworks, which is better than fixed-time control. The difference between subnetwork i and l is not obvious. In subnetwork j, hierarchical control yields a better performance than decentralized control, and a worse one than centralized control. The reason is that under these three control schemes, the traffic situation does not exceed the saturated situation too much, and along with the decrease of the traffic demand, the traffic situation becomes better. This can be verified in Fig. 10(d), 10(e) and 10(f). The number of vehicles in subnetwork i under hierarchical control is larger than under centralized control, and is less than under decentralized control, as shown in Fig. 10(d). It decreases after the 25th control step because of the decreasing traffic demand, and it then returns to the neighborhood of the critical point. In Fig. 10(e), hierarchical control keeps the number of vehicles in subnetwork j at a comparatively stable state after the traffic condition reaches the peak hour situation, while decentralized control leads to the congestion state with the increasing of number of vehicles and centralized control improves the traffic situation. In subnetwork l, hierarchical control and decentralized control both approach the performance of centralized control, as shown in Fig. 10(f). Moreover, for subnetwork i the average difference between the number of vehicles and the critical point under hierarchical control is 132 veh and

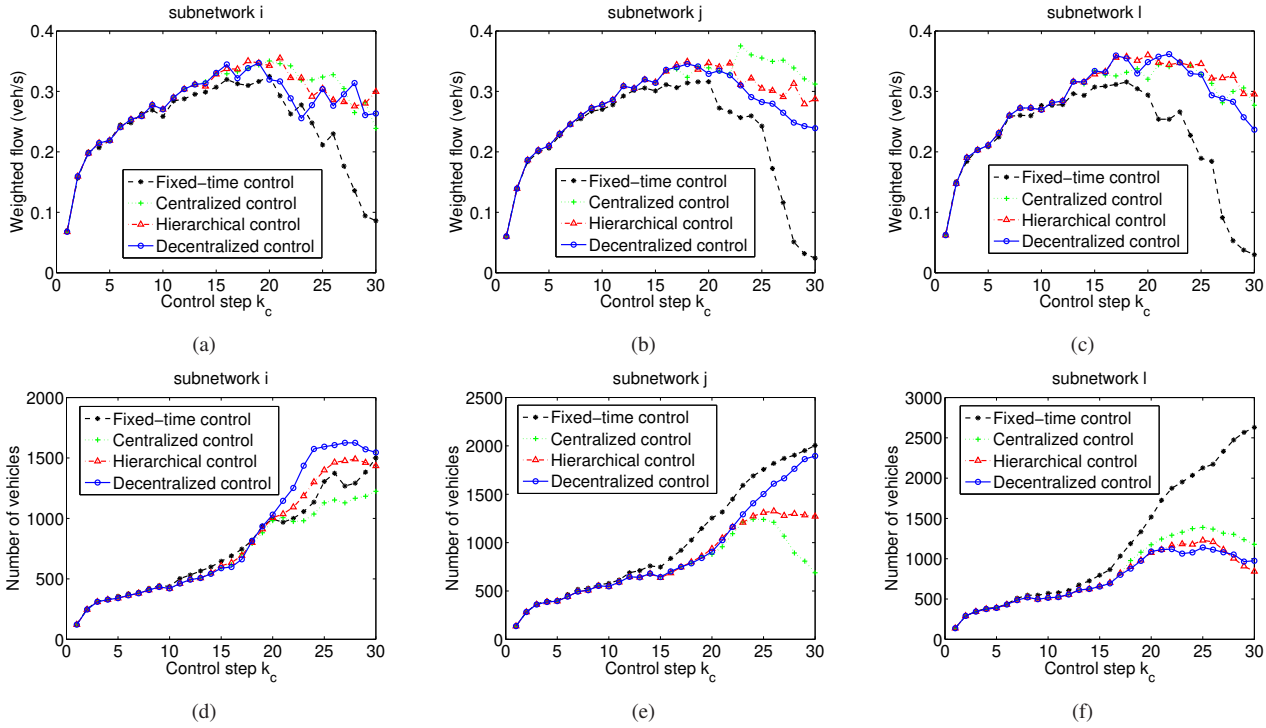

Fig. 10. Comparison for two performance indices of the four control strategies for three subnetworks in Scenario 2. (a), (b) and (c) Weighted average flow. (d), (e) and (f) Number of vehicles.

the number of vehicles under centralized control does not exceed the critical point, for subnetwork j, both the number of vehicles under hierarchical and centralized control do not exceed the critical point, and for subnetwork l, the average difference under hierarchical control is less than under centralized control. It seems that centralized control achieves a better performance than hierarchical control. However, note that in subnetwork j the number of vehicles decreases quickly with the decrease of traffic demand, while the traffic states are still in a relatively congested situation in the other subnetworks, especially in subnetwork l. In contrast, hierarchical control keeps the traffic state in each subnetwork in an appropriate situation, and along with the decrease of traffic demand from source nodes, the traffic states in the three subnetworks return from the congested situation synchronously. These results illustrate the fact that the coordination within hierarchical control plays an important role in balancing the traffic demands among subnetworks. It keeps the traffic state in each subnetwork at an appropriate situation. Therefore, the rate of change of number of vehicles at the end of the simulation is more balanced for hierarchical than centralized control.

All optimization problems are solved in a MATLAB 7.11 environment on a computer with a 3.20-GHz Intel Core (TM) I5 processor and 4-Gb RAM. In our case studies, there are two control variables at each signalized intersections. Moreover, the test traffic network has 34 signalized intersections and 154 two-way links, and the prediction horizon is 7. In the optimization problem of centralized control, there are $2\times 34\times 7=476$ control variables and $154\times 7+34\times 7+2\times 34\times 7=1792$ constraints. In the decentralized MPC strategy, the whole network is decomposed into three subnetworks. The largest subnetwork contains 12 signalized intersections and 55 links. Therefore, there are $2\times 12\times 7=168$ control variables and $55\times 7+12\times 7+2\times 12\times 7=637$ constraints in the corresponding optimization problem. In the hierarchical MPC strategy, each subnetwork is further decomposed into two subregions for the application of the multi-agent MPC approach. The largest subregion contains 6 signalized intersections and 29 links. There are $2\times 6\times 7=84$ control variables and $29\times 7+6\times 7+2\times 6\times 7=329$ constraints in the lower-level optimization problem of hierarchical control. Moreover, at the upper level, there are only $6\times 7=42$ control variables and $6\times 7+6\times 7=84$ constraints. In the case studies, we assume that there is no communication delays between the upper level and the lower level, and between the agents at the lower level. Since the scale of the optimization problem in hierarchical control is decreased, the CPU time for obtaining the optimal solutions is much less than that of the other approaches. The average and maximum CPU time spent for one run by the three control approaches is shown in Table III, which illustrates that the computational complexity of hierarchical control is significantly reduced. The average (maximum) CPU time represents the average (maximum) computation time used for solving the on-line optimization problems at each control time step. It should be noted that the SQP algorithm has been applied 5 times in each control step. Since the simulations are carried out on a single computer in our case studies, the SQP optimizations are run one by one to obtain the final solution. Therefore, the actual computation time of the multi-start technique of each control approach is 5 times of the average CPU time.

TABLE III
COMPARISON OF CPU TIME SPENT FOR THE THREE CONTROL STRATEGIES

Control strategy	Average CPU time (s)	Max CPU time (s)
Centralized control	797.2	1055.3
Decentralized control	149.6	193.0
Hierarchical control	80.7	105.8

VI. CONCLUSIONS AND FUTURE WORK

Network-wide traffic control plays an important role in mitigating and avoiding congestion in urban traffic networks. In this paper, based on a partition of the network, we have proposed a two-level hierarchical control scheme, the traffic demand balancing controller at the upper level together with the signal optimization controllers at the lower level to control a large-scale urban traffic network. The optimization problem at each level is formulated with a different traffic model and a different objective function. Through communication, the controllers work collaboratively to regulate the traffic flow and to guarantee a better performance of the whole network. Moreover, in order to reduce the computation time, a parallel distributed control scheme is introduced at the lower level to coordinate the subregion controllers, making them reach an agreement on their control decisions through negotiations. All optimization problems are embedded in an MPC scheme for real-time implementation. The simulation results under two different traffic demand scenarios show that the proposed hierarchical control approach can increase the weighted average traffic flow by keeping the number of vehicles approaching to the critical point of the MFD, and yield an efficient performance that is comparatively close to the results of centralized MPC. The results of the case studies illustrate the importance of the coordinating traffic demands in our approach compared with the decentralized control. Furthermore, it should be noted that the computation times required for solving the optimization problems in our approach are much lower than the other two MPC control methods.

In the future, we will explore how to increase the computation speed for solving the non-linear non-convex MPC optimization problem. Possible approaches to reduce the computation time, such as fast MPC [4], parallel computing and parameterized MPC [35], will also be investigated to further improve the computational efficiency and to make the proposed approach applicable in practice. In addition, other traffic performance objectives such as the L^2-norm and the L^∞-norm [36] could be taken into account in the optimization problem to make the traffic network more homogeneous. Several formal guarantees will also be investigated in the future. There are two ways to guarantee the recursive feasibility of the MPC controller. The first is to develop a robust MPC control approach, and the second is to introduce additional positive variables, such as soft constraints in optimization problems. The stability could be guaranteed by designing the corresponding Lyapunov function. In [37], they pave the way for the investigation of stability of non-convex optimization problem using PWA-based method.

ACKNOWLEDGMENT

This work is supported in part by the National Science Foundation of China (Grant No. 61433002, 61521063, 61374110, 61473288), the Beijing Natural Science Foundation (Grant No. 4142055), Chinese International Cooperation Project of National Science Committee (Grant No. 71361130012), and the NWO-NFSC project 'Multi-level predictive traffic control for large-scale urban networks' (629.001.011), which is partly financed by the Netherlands Organization for Scientific Research (NWO).

REFERENCES

[1] C. Diakaki, M. Papageorgiou, and K. Aboudolas, "A multivariable regulator approach to traffic-responsive network-wide signal control," Control Engineering Practice, vol. 10, no. 2, pp. 183-195, Feb. 2002.

[2] K. Aboudolas, M. Papageorgiou, and E. Kosmatopoulos, "Store-and-forward based methods for the signal control problem in large-scale congested urban road networks," Transportation Research Part C: Emerging Technologies, vol. 17, no. 2, pp. 163-174, Apr. 2009.

[3] K. Aboudolas, M. Papageorgiou, A. Kouvelas, and E. Kosmatopoulos, "A rolling-horizon quadratic-programming approach to the signal control problem in large-scale congested urban road networks," Transportation Research Part C: Emerging Technologies, vol. 18, no. 5, pp. 680-694, Oct. 2010.

[4] S. Lin, B. De Schutter, Y. Xi, and H. Hellendoorn, "Fast model predictive control for urban road networks via MILP," IEEE Transactions on Intelligent Transportation Systems, vol. 12, no. 3, pp. 846-856, Sep. 2011.

[5] S. Lin, B. De Schutter, Y. Xi, and H. Hellendoorn, "Efficient network-wide model-based predictive control for urban traffic networks," Transportation Research Part C: Emerging Technologies, vol. 24, pp. 122-140, Oct. 2012.

[6] N. H. Gartner, F. J. Pooran, and C. M. Andrews, "Implementation of the opac adaptive control strategy in a traffic signal network," in Proceedings of IEEE Conference on Intelligent Transportation Systems, Oakland, USA, 2001, pp. 195-200.

[7] P. Mirchandani, and L. Head, "A real-time traffic signal control system: architecture, algorithms, and analysis," Transportation Research Part C: Emerging Technologies, vol. 9, no. 6, pp. 415-432, Dec. 2001.

[8] L. B. de Oliveira, and E. Camponogara, "Multi-agent model predictive control of signaling split in urban traffic networks," Transportation Research Part C: Emerging Technologies, vol. 18, no. 1, pp. 120-139, Feb. 2010.

[9] L. Baskar, B. De Schutter, and J. Hellendoorn, "Hierarchical model-based predictive control for intelligent vehicle highway systems: Regional controllers," in Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems, Madeira Island, Portugal, 2010, pp. 249-254.

[10] Y. Wang, M. Papageorgiou, and A. Messmer, "A real-time freeway network traffic surveillance tool," IEEE Transactions on Control Systems Technology, vol. 14, no. 1, pp. 18-32, Jan. 2006.

[11] Q. Kong, L. Li, B. Yan, F. Zhu, and G. Xiong, "Developing parallel control and management for urban traffic systems," IEEE Intelligent Systems, vol. 28, no. 3, pp. 66-69, May. 2013.

[12] A. Hegyi, B. De Schutter, and J. Hellendoorn,"Model predictive control for optimal coordination of ramp metering and variable speed limits," Transportation Research Part C: Emerging Technologies, vol. 13, no. 3, pp. 185-209, Jun. 2005.

[13] N. Geroliminis, and C. F. Daganzo, "Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings," Transportation Research Part B: Methodological, vol. 42, no. 9, pp. 759-770, Nov. 2008.

[14] N. Geroliminis, and J. Sun, "Properties of a well-defined macroscopic fundamental diagram for urban traffic," Transportation Research Part B: Methodological, vol. 45, no. 3, pp. 605-617, Mar. 2011.

[15] A. Mazloumian, N. Geroliminis, and D. Helbing, "The spatial variability of vehicle densities as determinant of urban network capacity," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 368, no. 1928, pp. 4627-4647, Oct. 2010.

[16] N. Geroliminis, J. Haddad, and M. Ramezani, "Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: A model predictive approach," IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 1, pp. 348-359, Mar. 2013.

[17] K. Aboudolas, and N. Geroliminis, "Perimeter and boundary flow control in multi-reservoir heterogeneous networks," Transportation Research Part B: Methodological, vol. 55, pp. 265-281, Sep. 2013.
[18] J. Haddad, M. Ramezani, and N. Geroliminis, "Cooperative traffic control of a mixed network with two urban regions and a freeway," Transportation Research Part B: Methodological, vol. 54, no. 8, pp. 17-36, Aug. 2013.
[19] M. Hajiahmadi, J. Haddad, B. De Schutter, and N. Geroliminis, "Optimal hybrid perimeter and switching plans control for urban traffic networks," IEEE Transactions on Control Systems Technology, vol. 23, no. 2, pp. 464-478, Mar. 2015.
[20] S. Lin, and Q. Kong, "A model-based demand-balance control for complex urban traffic networks," in Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems, Qingdao, China, 2014, pp. 2900-2905.
[21] S. Lin, Q. Kong, and Q. Huang, "A simulation analysis on the existence of network traffic flow equilibria," IEEE Transactions on Intelligent Transportation Systems, vol. 15, no. 4, pp. 1706-1713, Aug. 2014.
[22] M. Keyvan-Ekbatani, A. Kouvelas, I. Papamichail, and M. Papageorgiou, "Exploiting the fundamental diagram of urban networks for feedback-based gating," Transportation Research Part B: Methodological, vol. 46, no. 10, pp. 1393-1403, Dec. 2012.
[23] S. Lin, T. Ling, and Y. Xi, "Model predictive control for largescale urban traffic networks with a multi-level hierarchy," in Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems, The Hague, The Netherlands, 2013, pp. 211-216.
[24] D. Q. Mayne and J. B. Rawlings, Model Predictive Control: Theory and Design, Madison, USA: Nob Hill Publishing, 2009.
[25] Y. Ji, and N. Geroliminis, "On the spatial partitioning of urban transportation networks," Transportation Research Part B: Methodological, vol. 46, no. 10, pp. 1639-1656, Dec. 2012.
[26] Z. Zhou, S. Lin, and Y. Xi, "A dynamic network partition method for heterogenous urban traffic networks," in Proceedings of the 15th International IEEE Conference on Intelligent Transportation Systems, Anchorage, USA, 2012, pp. 820-825.
[27] C. F. Daganzo, and N. Geroliminis, "An analytical approximation for the macroscopic fundamental diagram of urban traffic," Transportation Research Part B: Methodological, vol. 42, no. 9, pp. 771-781, Nov. 2008.
[28] D. P. Bertsekas, Constrained optimization and Lagrange multiplier methods, Computer Science and Applied Mathematics, Boston, USA: Academic Press, 1982.
[29] D. P. Bertsekas, and J. N. Tsitsiklis, Parallel and distributed computation: numerical methods, Englewood Cliffs, NJ, USA: Prentice-Hall, 1989.
[30] S. P. Boyd, and L. Vandenberghe, Convex optimization, Cambridge University Press, 2004.
[31] R. Negenborn, B. De Schutter, and J. Hellendoorn, "Multi-agent model predictive control for transportation networks: Serial versus parallel schemes," Engineering Applications of Artificial Intelligence, vol. 21, no. 3, pp. 353-366, Apr. 2008.
[32] G. Cohen. "Auxiliary problem principle and decomposition of optimization problems," Journal of optimization Theory and Applications, vol. 32, no. 3, pp. 277-305, Nov. 1980.
[33] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," Foundations and Trends® in Machine Learning, vol. 3, no. 1, pp. 1-122, Jan. 2011.
[34] Z. Zhou, B. De Schutter, S. Lin, and Y. Xi, "Multi-agent model-based predictive control for large-scale urban traffic networks using a serial scheme," IET Control Theory and Applications, vol. 9, no. 3, pp. 475-484, Feb. 2015.
[35] S. Zegeye, B. De Schutter, J. Hellendoorn, E. Breunesse, and A. Hegyi, "A predictive traffic controller for sustainable mobility using parameterized control policies," IEEE Transactions on Intelligent Transportation Systems, vol. 13, no. 3, pp. 1420-1429, Sept. 2012.
[36] S. Lin, Z. Zhou, and Y. Xi, "Model-Based Traffic Congestion Control in Urban Road Networks: Analysis of Performance Criteria," Transportation Research Record: Journal of the Transportation Research Board, no. 2390, pp. 112-120, Feb. 2013.
[37] M. Hajiahmadi, B. De Schutter, and H. Hellendoorn, "Design of stabilizing switching laws for mixed switching affine systems," IEEE Transactions on Automatic Control, to appear.