Network wide service level-oriented route guidance in road traffic networks

Traffic management in line with the policy objectives of the road authorities

TRAIL Research School, October 2012

Authors

Ir. Ramon Landman, Dr. Andreas Hegyi, Prof. dr. ir. Serge Hoogendoorn
Delft University of Technology, Faculty of Civil Engineering and Geosciences, Department of Transport & Planning, the Netherlands

Contents

Abstract

Service level control is a promising strategy to realize policy objectives within road networks like improving network production while taking road user interests, livability and safety into account. In this contribution a service level-oriented route guidance control approach is presented that is able to control service levels of routes on a network scale. By maintaining target service levels network performance degradation is prevented, redundant route capacity fully utilized and road user interests protected. By means of a simulation test case, the performance and functioning of the controller is compared with a Model Predictive Control based route guidance approach that realizes system optimal conditions and a user equilibrium feedback approach that realizes user optimal conditions. The results indicate that both the proposed and Model Predictive Control approach outperform the user equilibrium approach. However, the proposed approach approximates system optimal performance with a significantly lower computational demand and a better scalability with growing network sizes.

Keywords

Service level control; Network wide traffic management; Traffic management policy; Finite-state machine; Feedback control.

1 Introduction

Today's increasing adverse effects of congestion clearly indicate the need to apply traffic management on a network level to improve the network performance. This is a complex problem in which the traffic management policies of the road authorities need to be taken into account. Efficient traffic flows are obviously important, but aspects like road user interests, environment, livability and safety need to be considered too. To successfully realize traffic conditions that are in line with the policy objectives, a systematic control approach is necessary. Moreover, the system should produce control actions that are comprehensible for the authorities, because they are the ones responsible for the effects and consequences.

However, no control approaches are available yet that are able to integrally realize these different policy objectives in real time into practice. In this contribution therefore a service level-oriented route guidance approach is presented that is able to guide traffic such that network performance is improved while the other policy objectives remain respected. The approach is designed around the methodology that is used in the Netherlands to harmonize the interests of involved stakeholders with respect to operational network wide traffic management.

In previous work (Landman, 2011, 2012) we have shown that by means of service level-oriented route guidance control over two routes, it is possible to prevent phenomena that cause decreased network performance while also taking policy objectives other than network throughput into account. However, when maintaining service levels of important routes within a network by multiple actuators, it is well possible that the routes under control completely or partly overlap. Under these circumstances there will be interaction between the controllers, because the traffic that flows towards an overlapping route stretch can be manipulated from multiple directions. In this contribution we will show that service levels can be properly maintained within overlapping routes by multiple distributed actuators and that network performance is improved compared to a user equilibrium route guidance approach.

Not only the performance in terms of network throughput is important, also its computational demand and hence its scalability are essential with respect to real-time operationalization on a network level. Computational demand scales with the number of control signals that needs to be determined per control interval and the complexity of the control approach. Complex approaches can therefore pose limitations on their real-time applicability. The proposed approach will therefore be compared with such complex (i.e. a Model Predictive Control based approach) that is able to determine the optimal control signals given prevailing policy objectives. It turns out that it is possible after proper tuning to approximate system optimal behavior, however, against limited computational demand.

In the remainder of this section we discuss the background of this research and the requirements for control approaches to systematically improve network performance. In Section 2 the service level control methodology is elaborated including a description on how the control mechanism works for overlapping routes. The test case set-up and the controller settings are given in Section 3, followed by the results and discussion in Sections 4 and 5.

1.1 Developments in operational traffic management

To start with the realization of network-wide traffic management in practice, a method called `Sustainable Traffic Management' was developed in the Netherlands that harmonizes the different interests of involved stakeholders (Rijkswaterstaat, 2003). The output of the method is a common vision on the functioning of the network and all of its elements, expressed in terms of road priorities and corresponding desired service levels (van Kooten, 2011). In Figure 1 an example is given of such priority map for the city of Den Bosch in the Netherlands. The colors indicate the priorities given to the road stretches. Notice that the priorities of the elements decrease with increasing priority index (i.e. priority of 1 means most important).

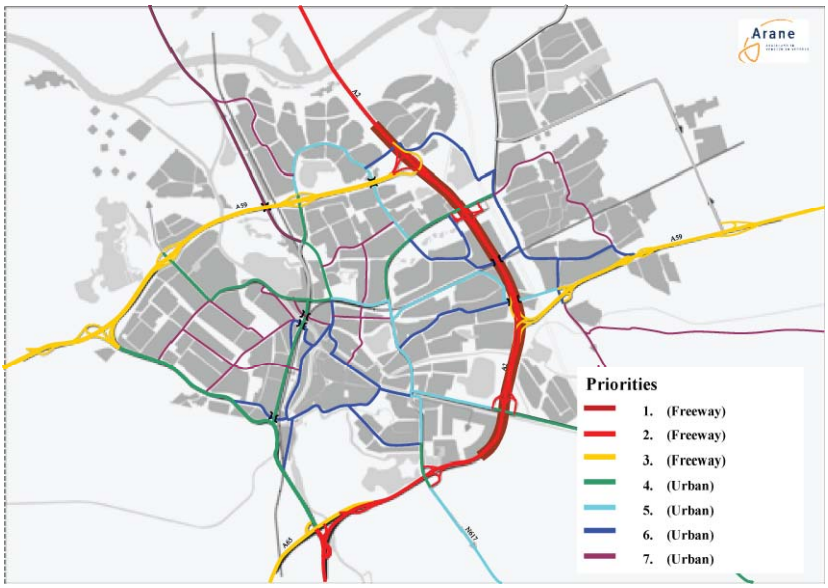

Figure 1: Priority map for the city of Den Bosch in the Netherlands indicating with different colors the priorities of the various network elements

The priority given to a road depends on aspects like the road's function, its average daily load, and its contribution to facilitating movements between important activity areas in the region. High capacity freeways are in this respect considered more important than rural roads and ring roads more important than arterials.

The service levels for the road stretches and routes are defined as performance ranges, indicated by an upper and lower bound of the traffic speed. The bounds determine the acceptable performance within a service level and they can differ over the various network elements. Hence, by equally degrading the service levels, different performance regimes can be established among the network elements in line with the objectives.

The priorities and service levels are thus the foundation upon which to decide where to guide or store traffic so that network performance degradation is prevented. Moreover, they are a compromise between on the one hand the harmonization process of the interests of involved stakeholders and on the other hand thorough traffic analysis to prevent phenomena that cause decreased network performance. In other

words, the bounds reflect objectives like increasing network performance, restrictions on the use of specific roads from a livability or safety point of view and functional criteria that define a maximal quality difference between the controlled routes (fairness).

One might however question if the priorities can be considered static, because network performance (i.e. the network outflow) is directly determined by the phenomena spill back and capacity drop[1]. Depending on the actual conditions these phenomena dynamically reduce the network outflow to a certain extent, and they need to be dealt with sagely to prevent decreased network performance. With respect to route guidance on a network level, spill back is the most important phenomenon to prevent. Its impact on the network performance is directly determined by the hindered amount of traffic that does not need to pass the bottleneck in combination with the bottlenecks strength (i.e. ratio between demand into the bottleneck and the bottleneck supply). Hence, priorities of the routes can change dynamically depending on the traffic flows that are or become blocked over time. The static priority map that results from the harmonization process between stakeholders is nevertheless a good starting point, because the routes' service level bounds can be chosen such that during oversaturated conditions the blocking of flows is realized in line with the order of the total impact on the network performance. This impact can be estimated based on empirical or simulation data beforehand or on-line.

The harmonization of interests with respect to the network functioning in the field of operational traffic management obviously indicates the need for a control system to maintain service levels over the network by means of coordinated network-wide Dynamic Traffic Management (DTM). In (Landman 2010) a first step towards a control framework was presented based on the strategy to protect the performance of important network elements at the cost of less important ones. Within such framework, route guidance can be considered as an important means to distribute traffic in line with the policy objectives over the network. A next step towards the operationalization is made by (Landman, 2011, 2012), presenting a control approach to degrade and recover the performance at two alternative routes by means of a finite-state machine. With respect to the applicability of the approach on a network level it then becomes important to investigate the interaction between distributed actuators that influence the service levels at the same locations.

1.2 Improving network performance

In literature many different control approaches can be found for applying dynamic route guidance. In this review the approaches are evaluated on their ability to systematically improve the network performance and to realize policy objectives in practice.

A feedback mechanism is required to deal with the many uncertainties in traffic demand and infrastructure supply. Most applied automated route guidance systems are therefore of the reactive (Mammar, 1996; Pavlis, 1999; Minciardi, 2001) or

[1] The capacity of a freeway road stretch drops with the onset of congestion because the flow out of the head of the queue is less than the maximal achievable flow in a free flow regime and queues that block upstream infrastructure (intersections and off ramps) cause hindrance to traffic that does not need to pass the bottleneck.

predictive (Messmer, 1998; Wang, 2003) feedback control type. Their objective is to equalize (arrival, instantaneous or departure) travel times over routes, and hence support the road user in making optimal routing decisions. However, in (Landman, 2011, 2012) it is shown that user equilibrium (UE) conditions might severely decrease network performance, especially when guiding traffic over routes of different priority. Moreover, reactive feedback (route guidance) systems are known to be vulnerable to system oscillations when the impact of given control signals is delayed. This is the case when the location of the actuator and desired impact are different, and if arrival or instantaneous travel times are used as feedback information (Wahle, 2000; Davis, 2010; Hoogendoorn1997). On a network scale (e.g. when applying DTM measures to solve traffic problems elsewhere in the network) the impact of signals can be strongly delayed and therefore even counterproductive. Even though predicted state estimates may prevent these instabilities (Wang 2001, 2003), user equilibrium feedback is not suited to systematically improve network performance and operationalize policy objectives.

Another important requirement for improving network performance is to gain insight into the effects of control signals on future traffic conditions. Optimization-based approaches attempt to optimize a network performance measure by applying a traffic flow model in an iterative optimization procedure. Two optimization-based approaches can be distinguished: Optimal Control (OC) and Model Predictive Control (MPC). In OC (Hoogendoorn, 1997; Papageorgiou 1990) the signals are optimized over some predefined period based on an initial state and expected demands. Due to the lack of a feedback loop, unexpected disturbances can make the previously optimized signals suboptimal. OC also has a large computational demand that makes application on large-scale networks not feasible in real time. This feedforward technique is therefore not suited to be applied in operational traffic management.

In MPC the optimization of control signals is done in a rolling horizon that entails a feedback mechanism, making it a potential alternative to the proposed approach. However, due to its complexity and computational demand it is predominantly used off-line to analyze for instance System Optimal (SO) and User Equilibrium (UE) solutions for large-scale networks (Peeta, 1995; Messmer, 1995), to realize SO route guidance signals by optimizing marginal costs (Zuurbier, 2006, 2010), or to determine optimal route guidance in combination with other DTM measures within urban and freeway networks (Karimi, 2004; van den Berg, 2004). Within the test case it can nevertheless be used since policy objectives can be incorporated in the objective function that is optimized.

2 Control Approach

A finite-state machine in combination with feedback control laws is used to maintain the desired the service levels within the controlled routes. In this section the functioning of the finite-state machine is described for guiding traffic over a single route set consisting of two routes and for multiple route sets where routes overlap.

2.1 Service level control for two routes

A finite-state machine is designed to dynamically determine the target service levels to realize within routes. This is done by a stepwise degradation and recovery scheme that ensures that the performance within the controlled routes can be stabilized such that spill back is prevented in the route where it most strongly threatens the overall network performance. Based on the prevailing traffic conditions with respect to the predefined service levels, the finite-state machine decides on which route the target service level is maintained, so that the other route is allowed to further degrade or recover during respectively over- and undersaturated conditions. Oversaturated means that the traffic demand for both routes is larger than their joint capacities, resulting in increasing congestion and decreasing service levels. If the demand for both routes is smaller than this joint capacity, routes are said to be undersaturated (even though congestion can still be present), resulting in performance recovery. In the remainder of the contribution, the route that generally carries the largest flows or that has the largest supply is considered the main route with index $r = 1$ and the corresponding alternative then receives index $r = 2$.

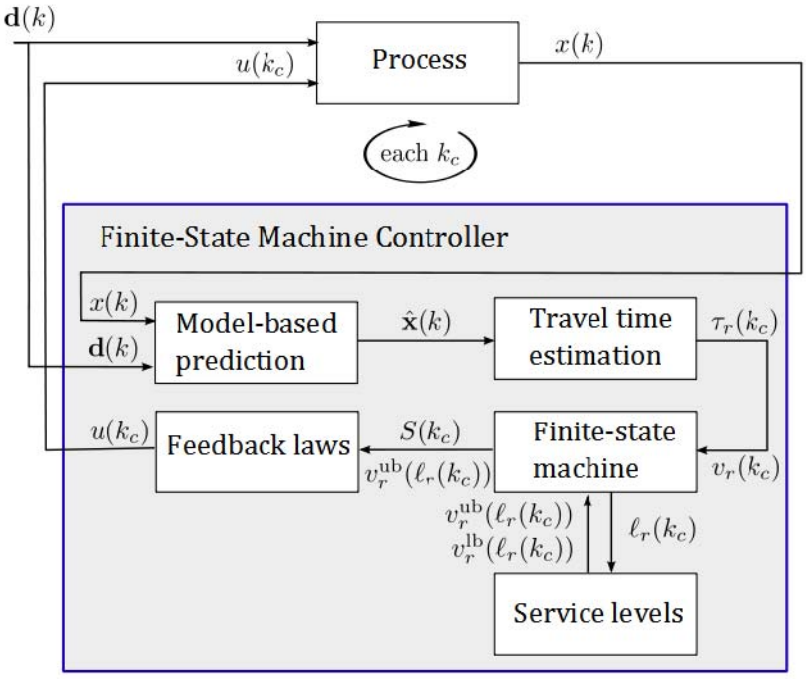

Figure 2: The finite-state machine control loop

In Figure 2 the control loop is shown. Notice, that `Process' in fact is not a model but the real traffic process. The simulation time step counter k and the control time step counter k_c indicate time instants kT and $k_c T_c$, with T and T_c respectively the simulation and controller time step sizes. We assume that $T_c = TM$, with M being an integer. The simulation process is modeled by the discrete-time system f:

$$x(k+1) = f(x(k), u(k_c), d(k)) \tag{1}$$

with $Mk_c \le k \le M(k_c+1)$, $x(k)$ the state vector of the system (e.g. flows, speeds, densities over links) at simulation step k, $u(k_c)$ the control input at controller time step k_c (e.g. controlled split fractions) and $d(k)$ the disturbance vector (e.g. demands) at simulation step k.

When the controller is activated, the state vector $x(k)$ is the initial state for a model-based prediction that is used to define the future traffic conditions $\hat{\mathbf{x}}(k)$ over some prediction horizon. Based on this prediction, the departure travel time $\tau_r(k_c)$ for each route r is determined[2].

The travel times indicate the current performance of each route, and they are easily translated to an average travel speed $v_r(k_c)$ in km/h in combination with the route length. Based on these performance indications and the predefined service levels, the finite-state machine decides upon which feedback algorithm to activate to compose the control signal.

Table 1: Service levels $l_r(k_c)$ with their upper $v_r^{ub}(l_r(k_c))$ and lower bounds $v_r^{lb}(l_r(k_c))$ expressed in km/h

Service Level	Main route		Alternative	
$l_r(k_c)$	$v_1^{\mathrm{ub}}(l_1(k_c))$	$v_1^{\mathrm{lb}}(l_1(k_c))$	$v_2^{\mathrm{ub}}(l_2(k_c))$	$v_2^{\mathrm{lb}}(l_2(k_c))$
1	80	60	80	50
2	60	40	50	30
3	40	20	30	20
4	20	10	20	10
5	10	0	10	0

The service levels are expressed in terms of travel time or speed, and in Table 1 an example is given in terms of speed, because this gives a generic performance description that is not dependent on route lengths. With respect to the implementation, the service levels are always translated into travel times, because this prevents unrealistic and unfair travel time differences between route alternatives from being realized and maintained[3].

For each route, every service level $l_r(k_c)$ is determined by an upper bound $v_r^{\mathrm{ub}}(l_r(k_c))$ and lower bound $v_r^{\mathrm{lb}}(l_r(k_c))$. Notice from the table that the bounds of the same service level can be different for the different routes, and that the level indexes increase when the performance degrades.

[2] The length of the prediction horizon is defined by the maximum travel time through each route. The travel times are determined by a trajectory-based method applied on the predicted speed profiles of the routes (van Lint, 2010). The finite-state machine, however, also allows for the use of arrival or instantaneous travel times.

[3] Due to the relation $\tau_r = L_r / v_r$, with v_r the speed, L_r the length and τ_r the travel time of route r, small variations in low speeds result in much larger travel time differences than small variations in high speeds.

At the moment the finite-state machine is activated, it updates its state $S(k_c) \in \{1,2\}$ and the active service levels $l_r(k_c)$ based on a comparison between the actual route performance $v_r(k_c)$ and the active service level bounds $v_r^{\mathrm{ub}}(l_r(k_c-1))$ and $v_r^{\mathrm{lb}}(l_r(k_c-1))$ of the previous control interval from the route of which the performance is not kept constant.

The updated state $S(k_c)$ is subsequently used to select and execute the corresponding feedback algorithm with the setpoint $v_r^{\mathrm{ub}}(l_r(k_c))$ to determine the control signal. In state $S(k_c)=1$ the performance of the main route is kept constant by using $v_1^{\mathrm{ub}}(l_1(k_c))$ as setpoint allowing the alternative to degrade or recover and in $S(k_c)=2$ the applied setpoint is $v_2^{\mathrm{ub}}(l_2(k_c))$ allowing the main route to degrade or recover.

In Figure 3 the finite-state machine is shown. The outer loop over the finite-state machine states (red) is followed during the degradation process and the inner loop (green) during the recovery process. If the performance of the route that is allowed to degrade or recover remains within its service level bounds, no state transition is triggered and the active service levels remain the same. This is indicated by the triggers and actions (blue) within each state.

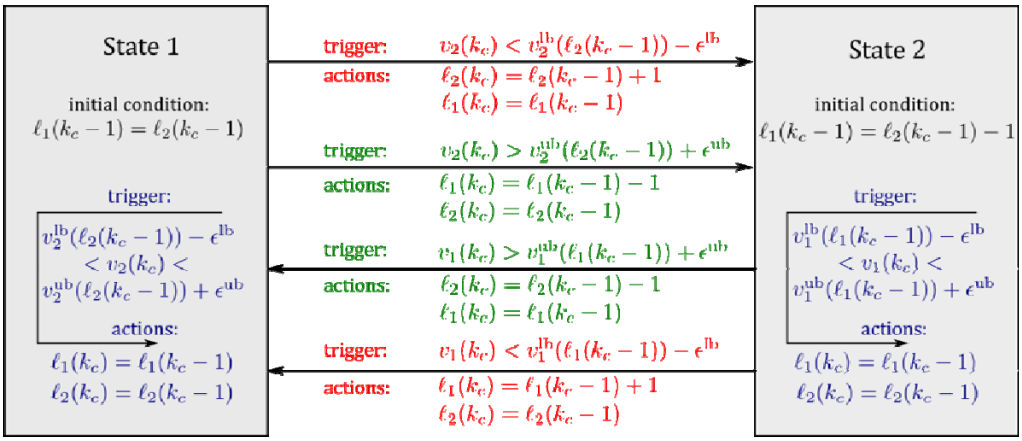

Figure 3: The finite-state machine with in the squares the different states of the system, and on the arrows the triggers to make a state transition and the corresponding action of switching the service level. In state $S(k_c)=1$ the service levels of main route and alternative are equal, and in state $S(k_c)=2$ the service level of the alternative is one index step higher than that of the main route.

To prevent frequent switching, an extra threshold is added to the bounds that trigger a state transition. This threshold is a constant value μ defined in terms of travel time. However, since the process description is in terms of speed, μ is translated into the terms $\varepsilon^{\mathrm{lb}}$ and $\varepsilon^{\mathrm{ub}}$ expressing the threshold as a function of the route length L_r, the considered reference value $v_r^{\mathrm{lb}}(l_r(k_c))$ or $v_r^{\mathrm{ub}}(l_r(k_c))$, and the defined travel time

difference μ. The upper and lower bounds become respectively $v_r^{\text{ub}}(l_r(k_c)) + \varepsilon^{\text{ub}}$ and $v_r^{\text{ub}}(l_r(k_c)) - \varepsilon^{\text{lb}}$.

The applied feedback control laws are given in (2). They determine the desired split fraction $\beta_n^d(k_c)$ (this is the control signal $u(k_c)$ in Figure 2 for the controllable traffic flow at the node n directly downstream the VSM towards destination d in control interval k_c. The desired split fraction $\beta_n^d(k_c)$ is a function of the previously applied split fraction $\hat{\beta}_n^d(k_c - 1)$, a feedback gain α, the current route performance in terms of travel time $\tau_r(k_c)$, and the setpoint $\tau_r^{\text{ub}}(l_r(k_c))$ from the service level table in terms of travel time.

$$\beta_n^d(k_c) = \begin{cases} \hat{\beta}_n^d(k_c - 1) + \alpha\left(\tau_1(k_c) - \tau_1^{\text{ub}}\left(l_1(k_c)\right)\right), & S(k_c) = 1 \\ \hat{\beta}_n^d(k_c - 1) - \alpha\left(\tau_2(k_c) - \tau_2^{\text{ub}}\left(l_2(k_c)\right)\right), & S(k_c) = 2 \end{cases} \tag{2}$$

It has to satisfy $0 \le \beta_n^d(k_c) \le 1$ and therefore might need to be truncated by $\hat{\beta}_n^d(k_c) = \min(\max(0, \beta_n^d(k_c)), 1)$. The realized split fraction towards the main route $\tilde{\beta}_n^d(k_c)$, however, depends on the compliance (driver response) γ of the controlled flow, and the nominal fraction (default behavior) towards the main route $\beta_n^{\text{N},d}(k_c)$. The implemented split fraction at time step k_c towards the main route then becomes

$$\tilde{\beta}_n^d(k_c) = (1 - \gamma)\beta_n^{\text{N},d} + \gamma\hat{\beta}_n^d(k_c) \tag{3}$$

and towards the alternative

$$\tilde{\beta}_n^{\text{'}d}(k_c) = 1 - \tilde{\beta}_n^d(k_c) \tag{4}$$

2.2 Service level control for overlapping routes

When applying the approach for multiple route sets, route sets can share overlap with each other by their main or alternative route, and the number of possible overlap combinations is then determined by the number of involved sets. However, a simple interaction mechanism ensures the utilization of redundant capacity and stepwise performance degradation and recovery within the routes. It is assumed that all routes $r \in \{1, 2\}$ within the considered route sets $s \in S$ initially perform within their first service level $l_{r,s}(0) = 1$.

At the moment a bottleneck becomes active within a route stretch, first redundant capacity is utilized from the sets $z \in Z$ that are directly involved with $Z \subset S$. If the problem concerns the main route of set z, the target service level $v_{1,z}^{\text{ub}}(l_{1,z}(k_c))$ is maintained by sending traffic to the corresponding alternative to either use its redundant capacity or degrade it until its first service level lower bound $v_{2,z}^{\text{lb}}(l_{2,z}(k_c))$.

In case the bottleneck concerns the alternative, the performance is allowed to degrade to $v_{2,z}^{\mathrm{lb}}(l_{2,z}(k_c))$, its service level index subsequently increased $l_{2,z}(k_c) = l_{2,z}(k_c - 1) + 1$, and the corresponding upper bound $v_{2,z}^{\mathrm{ub}}(l_{2,z}(k_c))$ maintained. Traffic is then sent back to the main route to use its redundant capacity or to degrade its performance to the active service level lower bound $v_{1,z}^{\mathrm{lb}}(l_{1,z}(k_c))$.

At route sets $y \in Y$ with $Y \subset S$ that are not directly influenced by the bottleneck, redundant capacity is used as soon as the performance degrades within the route stretches that overlap. Again, traffic is directly sent towards the alternative within set y if the stretch belongs to a main route by maintaining $v_{1,y}^{\mathrm{ub}}(l_{1,y}(k_c))$. The controller of set y basically turns the capacity over to the surplus of traffic from set z with which it overlaps. The same reasoning holds if the overlap is realized by the alternative of set y, however, after the performance degraded to its first service level lower bound $v_{2,y}^{\mathrm{lb}}(l_{2,y}(k_c))$.

During oversaturated conditions, this mechanism realizes stepwise degradation and recovery within the routes. The reasoning is then in terms of storage space with respect to the actual performance and the active service level bounds. However, the controllable flow in combination with its compliance should be large enough to completely relieve bottlenecks within the main and alternative route.

3 Test case

By means of the test case the interaction between multiple route guidance actuators controlled by the proposed approach (maintaining service levels within overlapping routes) is illustrated and compared to a system optimal and user optimal approach.

The prevailing policy objectives are to maximally improve network throughput, but to keep congestion within the underlying road network as long as possible. Moreover, a minimal travel time difference of 3 minutes over the routes is maintained as long as possible, but when congestion spill back starts to threaten flows that do not move through the bottleneck this minimal travel time difference is relaxed.

In the remainder of this section, the applied traffic flow model, the performance indicators and the set-ups of the test case, the finite-state machine and the MPC controller are discussed. Further, special attention is given to the chosen service level bounds to realize the above mentioned policy objectives.

3.1 Applied traffic flow model

The macroscopic multi-class cell-based traffic flow model Fastlane (van Lint, 2008) has been used for the process simulation, the state predictions of the finite-state-machine and the optimization procedure within the MPC controller. The main advantage of Fastlane is that it correctly models the build up and solving of congestion including the negative effects of the blocking back phenomenon. The flows are modelled destination dependent by means of split fraction definitions at the

nodes. This enables correct manipulation and propagation of traffic between specific origin-destination pairs by means of route guidance.

3.2 Performance indicators

The different control methodologies are evaluated based on the network performance indicator: the total time that vehicles have spent in the network (TTS). The time spent by $N(k)$ vehicles in one time step is $TN(k)$ and the total time that the vehicles spend in the network over a period $k = \{0,1,...,K\}$ with K the total number of simulation time steps becomes

$$J_{TTS} = \zeta_1 T \sum_{k=1}^{K} \sum_{m \in M} \sum_{c \in C_m} \rho_{m,c}(k) \lambda_{m,c} \tag{5}$$

with $\rho_{m,c}(k)$ the vehicle densities (in veh/km) over the cells $c \in C_m$ of all network links $m \in M$, $\lambda_{m,c}$ the corresponding cell lengths (in km) and ζ_1 the functions weight factor.
With respect to the MPC objective function this definition is extended with the requirement to limited travel time differences over the complete prediction horizon

$$J_{MPC} = J_{TTS} + \zeta_2 \sum_{s \in S} \max(0, | \tau_{1,s} - \tau_{2,s} |) - \Delta \tau^{\max} \tag{6}$$

with $\tau_{r,s}$ a summation of the travel time differences between main route and alternative of route set $s \in S$ (determined every minute over the complete prediction horizon), and ζ_2 the functions weight factor. To conclude, realized travel times $\tau_r(k)$ on the routes are evaluated in combination with the applied control signals $u(k_c)$.

3.3 Test case set-up

The network for the test case given in Figure 4 is inspired by the situation at the northern part of the city Den Bosch in the Netherlands (from the priority map in Figure 1). There are two route sets of which the main routes overlap. All routes are can be used to guide traffic from the freeway to a large event area within the urban environment and vice versa. However, the focus in this case is on the first situation.

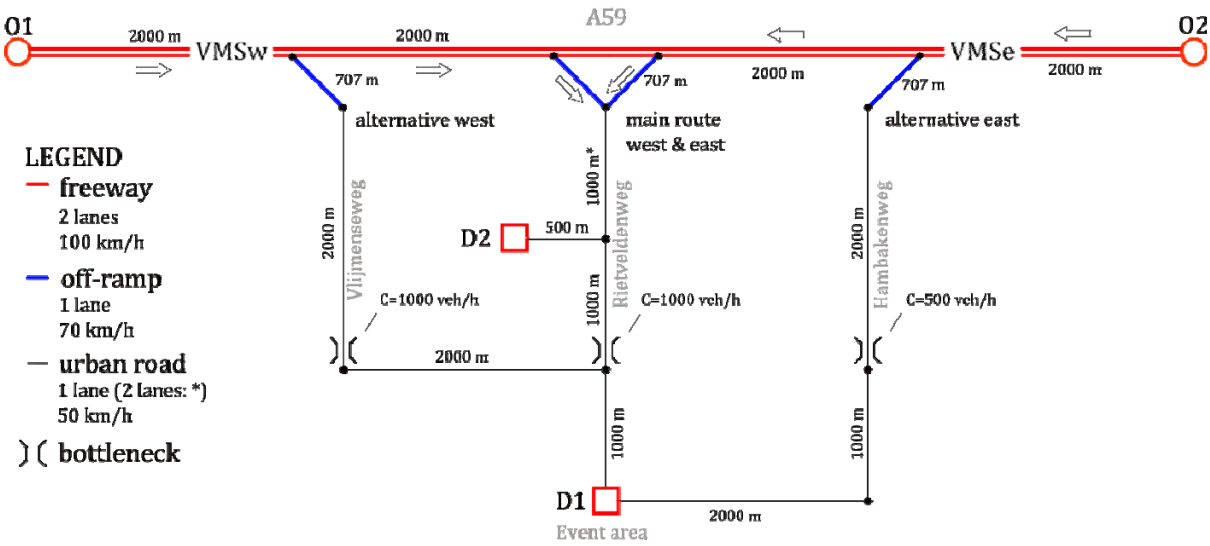

Figure 4: Network test case of the northern part of the city Den Bosch

The Variable Message Signs to distribute traffic are located in the west and east. The controlled traffic moves from origins O_1 and O_2 towards destinations D_1 in the south and the flow rates are interpolated from Table 2 over a 4-hour simulation period. From both directions, destination D_1 can be reached by a main route in the centre of the network, and alternatives at respectively the west and east side. The non-controllable flows move from origins O_1 and O_2 to destination D_2 and remain constant over the simulation period at respectively 500 and 250 vehicles per hour. Within each route a bottleneck is located with limited capacity (e.g. representing an intersection) to realize congestion. The compliance rate γ of traffic to a given advice is assumed to be 100% and the nominal split fraction $\beta_n^{N,d}$ at the nodes n downstream both actuators towards destination D_1 over the main routes is 90%.

Table 2: Demand pattern test case

Time	(hh:mm)	8:00	8:30	9:00	9:30	10:00	10:30	11:00	11:30	12:00
Flow O_1-D_1	(veh/h)	0	1000	1000	1000	1000	1250	1250	0	0
Flow O_2-D_2	(veh/h)	0	1000	1000	1500	1500	1500	1500	0	0

These settings result in the following scenario. Until 9:00AM, the demand towards D_1 is still smaller than the total route capacity. However, the initial demand for the main route severely exceeds its bottleneck capacity. From 9:00 to 9:30AM the demand from the O_2 increases and the total demand then equals the total route capacity. From 10:00AM to 10:30AM the demand from O_1 is increased which causes both route sets to become oversaturated.

3.4 Finite-state machine set-up

With respect to the finite-state machine, the applied service levels are given in Table 3. First the critical spill back conditions are mapped to the average condition in terms of travel time (based on empirical or simulation data). In this case spill back within the main routes block the flow towards D_2 at a travel time of 600 seconds. The off ramps within the main route, the alternative west and alternative east are respectively reached at travel times of 1000, 1000 and 1400 seconds.

These are important values in the service level table, together with the routes their free travel times and the desired degradation step size (i.e. in our case 100 seconds resulting in a maximal travel time difference of 3 minutes). The free travel times over the routes are the upper bounds of the first service level, and the lower bounds are acquired by adding the desired degradation step size. The other service levels bounds follow naturally. However, when filling in the bounds, the critical values with respect to spill back are approached, like the 3th service level upper bound of the main route (i.e. approaching 600 seconds). To prevent spill back from blocking the flow to D_2 the 3th service level upper bound of the main route is maintained and the alternatives are allowed to degrade till congestion reaches the off ramps at corresponding critical travel time values (i.e. the 3th service level lower bounds). Notice that the desired degradation step is relaxed with respect to the network performance.

If oversaturated conditions last, the congestion will be stabilized at the off ramps of the alternatives, allowing the main routes to degrade till their 3th service level lower bounds. Hence, it is accepted that the turning direction towards D_2 becomes blocked in order to prevent congestion spill back to the freeway on the alternatives. Notice, that when the main routes are subsequently degraded, congestion can no longer be prevented on the freeway network. From there on the routes are again degraded with the desired step size of 100 seconds.

Table 3: The odd columns of the service level table indicate the upper bounds $v_{r,s}^{\mathrm{ub}}(l_{r,s}(k_c))$ of the main routes and alternatives and the even columns their lower bounds $v_{r,s}^{\mathrm{lb}}(l_{r,s}(k_c))$ in terms of travel time (s). Notice that r represents the route index and s the corresponding route set index

S.L.	Main route		Alternative		Main route		Alternative	
$l_r(k_c)$	$v_{1,1}^{\mathrm{ub}}(l_{1,1}(k_c))$	$v_{1,1}^{\mathrm{lb}}(l_{1,1}(k_c))$	$v_{2,1}^{\mathrm{ub}}(l_{2,1}(k_c))$	$v_{2,1}^{\mathrm{lb}}(l_{2,1}(k_c))$	$v_{1,2}^{\mathrm{ub}}(l_{1,2}(k_c))$	$v_{1,2}^{\mathrm{lb}}(l_{1,2}(k_c))$	$v_{2,2}^{\mathrm{ub}}(l_{2,2}(k_c))$	$v_{2,2}^{\mathrm{lb}}(l_{2,2}(k_c))$
1	396	490	468	570	396	490	468	570
2	490	590	570	670	490	590	570	670
3	**590**	**1000**	670	**1000**	**590**	**1000**	670	**1400**
4	1000	1100	1000	1100	1000	1100	1400	1500
5	...	...	...	...	...	...	...	...

To conclude, the threshold μ is chosen as 10 seconds and the default feedback gain α is chosen 0.001.

3.5 Model Predictive Control and controller setup

A MPC scheme is used to solve the problem of realizing system optimal route guidance. At each time step k_c the optimal control signals $\mathbf{u}^*(k_c)$ are computed (by numerical optimization) over a prediction horizon N_p. A control horizon N_c ($< N_\mathrm{p}$) is selected to reduce the number of variables for optimization, and improve the stability of the system. In the optimization procedure, a model is used to evaluate the system performance over the prediction horizon based on the current state of the system $x(k)$, the expected disturbances $\mathbf{d}(k)$, and some planned control signals $\mathbf{u}(k_c)$. The corresponding performance of the system (e.g. the total time spent by vehicles in the system) is then evaluated by an objective function $J(\hat{\mathbf{x}}(k),\mathbf{u}(k_c))$ based on the evolution of the states $\hat{\mathbf{x}}(k)$ and the control signals $\mathbf{u}(k_c)$ within the prediction horizon. The optimization procedure minimizes the objective function's value by means of a suitable optimization algorithm. From the resulting optimal signals only the first sample $u^*(k_c)$ is applied to the process. In the next control time step (k_c+1), a new optimization is performed (with a prediction horizon that is shifted one control time step ahead) and of the resulting control signal again only the first sample is applied, and so on. This scheme, called rolling horizon, allows for updating the state from measurements in every iteration step. For more information on MPC see (Hegyi, 2004) and the references therein. To conclude, the control signals that are activated in the traffic process need to be translated to the actual split fraction of the controllable flow that responds to the advice given some assumed compliance rate γ. The same procedure is applied as given in (3) and (4).

When applying MPC, it is very important to determine the correct settings for the prediction horizon N_p, the number of variable control signals within the control horizon N_c, and of course the size of the parameter M that directly determines the size of the control interval $T_c = MT$ for a given simulation time step size T. The main rule for tuning N_p is that the prediction horizon should be long enough to cover the important system dynamics to find optimal conditions. However, in this case we also want the MPC control trajectory to be interpretable. A 60-minute prediction horizon of 10 control intervals (M=100, T=3.6 seconds) and 4 variable control signals per actuator are sufficient to realize system optimal conditions by an interpretable control trajectory. The computational demand is analyzed by increasing the control horizon per actuator stepwise from 1 to 7.

4 Results

In this section the control signals are evaluated in relation to the resulting travel times on the main routes and their alternatives. For the finite-state machine approach, a process description is given by the graph of feedback gain 0.001 indicated by the black lines in Figure 5A, B, C and D. Then, some remarks are made about the consequences of overshoot and oscillation of control signals due to the size of a feedback gain. Finally, the results are compared with the MPC based approach and

the user equilibrium feedback approach including a short reflection on the applicability in practice.

4.1 Finite-state machine approach

From 8:00 to 8:30AM traffic from O_1 and O_2 activates the bottleneck on the main routes while there is enough redundant capacity on the alternatives. Figure 5A-I and B-I show that both controllers directly start redirecting traffic from the main routes to the alternatives to protect the main routes' performance. Notice that from 8:30AM on only the alternative on the west has redundant capacity left.

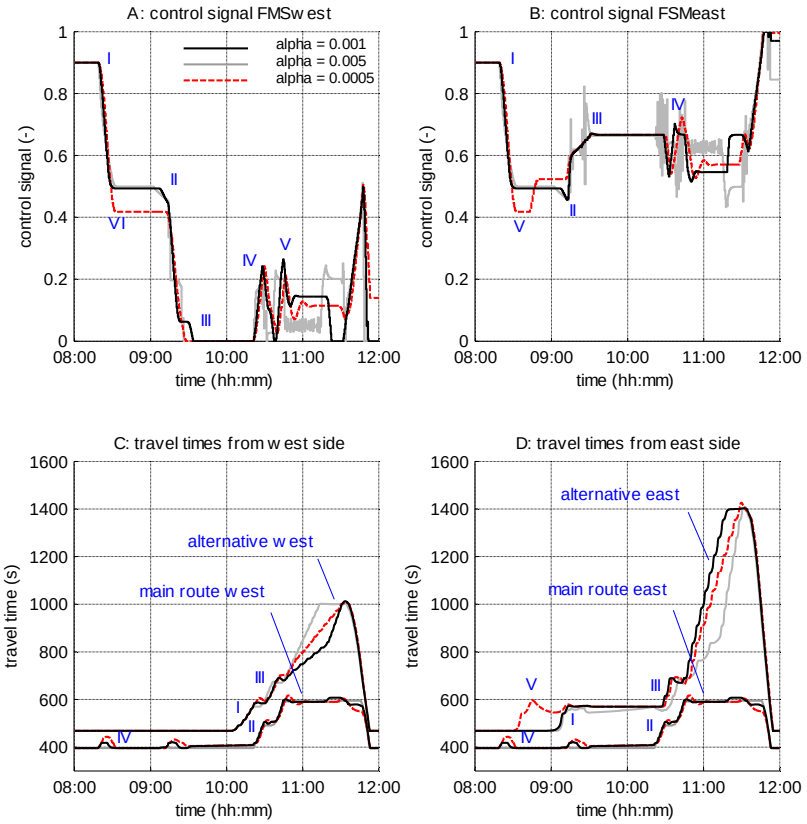

Figure 5: Control signals and travel times finite-state machine approach. Notice that a control signal of 1 means that all traffic is sent to the main route and 0 means that all traffic is sent to the alternative

From 9:00 to 9:30AM the traffic flow from the east increases with 500 vehicles per hour. Figure 5D-I shows that the extra traffic directly causes the travel time to increase at the alternative on the east until the lower bound of its first service level. Traffic is subsequently sent back to the main route until the inflow of the alternative is equal to the bottleneck capacity as can be seen in Figure 5B-II/III. In the mean time, the controller on the west side keeps the performance of both main routes constant by redirecting all traffic from the west to its alternative as can be seen in Figure 5A-II/III. It is this mechanism that realizes full utilization of available capacity over the routes.

From 10:00AM on, the flow from the west towards D_1 increases without any redundant capacity left. As can be seen in Figure 5A-III and C-I, all traffic from the west remains guided to the alternative in the west until the lower bound is reached from its first service level. Both alternatives are now degraded to their second service level, and from Figure 5A-IV and C-II we can see that traffic from the west is steered back towards the main route to degrade its performance till the lower bound of its first service level is reached. Since the alternative on the east is already in its second service level, also the main route east is accepted to degrade as can be seen in Figure 5D-II. From then on, Figure 5A-V and B-IV show that the signals start fluctuating to realize the desired stepwise decrease of the route performances, starting with both alternatives (Figure 5C-III and D-III). Finally, Figure 5C and D clearly show that the performance between main routes and their alternatives is degraded stepwise, including the large degradation of the alternatives to prevent spill back from blocking the turning direction in the main routes towards D_2.

4.2 Discussion on tuning the finite-state machines

Without having specified a complex coordination layer, both finite-state machines guide traffic such that redundant capacity is used during undersaturated conditions and routes degrade stepwise during oversaturated conditions. The small queue that is maintained at the alternative in the east to utilize the redundant capacity in the west can be considered reasonable, since the flow from the east caused the need to use redundant capacity elsewhere in the network.

The size of the feedback gain determines how well the controller deals with demand fluctuations (given demand and supply characteristics). Feedback gains that are too small (see α = 0.0005 by red dashed line in Figure 5) could realize overshoot that may trigger unnecessary and undesired congestion (increased travel time) when the controllers are not able to adequately reroute traffic. Hence, at the time the control signal is large enough to realize the required distribution, there is still congestion on the main route as can be seen in Figure 5C-IV and D-IV which causes the controllers to send even more traffic to the alternatives indicated at Figure 5A-VI and B-V. This is no problem for alternatives with sufficient redundant capacity as can be seen in Figure 5C around 8:30AM by the constant travel time at alternative west. However, on alternatives with limited capacity, travel times will instantly grow as can be seen in Figure 5D-V. In this case, the overshoot also caused traffic from the west to use the alternative more than needed, leaving space on the main route that is later used partly by the controller in the east to correct for the overshoot on the alternative east when maintaining its lower bound.

Feedback gains that are too large (see α = 0.005 by gray continuous line in Figure 5) cause the signal to oscillate as can be seen in Figure 5A and B. However, even though signal oscillations are undesired from an application point of view, they do realize the desired system behavior because the controllers deal adequately with situational changes. The proper size of the gains is strongly related to the size of the control intervals and the variations in the demand patterns, hence they are situation specific. The smaller the control intervals the smaller the gain can be due frequent corrections of the control signal, and the larger the demand variations the larger the gain must be to adequately respond on travel time differences to find a new equilibrium state.

4.3 MPC based approach

The important difference between MPC and the proposed method is that it anticipates on future traffic conditions and their effect on the network performance. The control signals are instantly adjusted instead of gradually achieved and chosen such that the objectives are exactly met. The controller first lets the bottleneck on the main route to become saturated as can be seen in Figure 6A-I and B-I, before sending traffic to the alternatives. In first instance all traffic from the west is sent to the alternative, however, Figure 6A-II shows that part is sent back to keep on utilizing the main routes full capacity. The main routes are the shortest and therefore using their full capacity positively affects the network performance. During oversaturation it further does not matter where the queues are located as long as redundant capacity on the alternatives is used and the flow to D_2 not hindered. Figure 6C and D show that congestion is kept limited at the main routes but accepted to grow on the alternatives. However, to prevent congestion spill back to the freeway, traffic from the alternative west is partly sent back to the main route (see Figure 6A-IV), compensated by sending (slightly more) traffic from the eastern main route to its alternative (see Figure 6B-II). Congestion quickly grows on the alternatives until it reaches the off ramps of the freeway (see Figure 6C-I and D-I). In the mean time the congestion within the main route dissolves, however, the travel time patterns indicate that it does not become underutilized (see Figure 6C-II and D-II). The reason for the controller to accept the direction towards D_2 to become shortly blocked at Figure 6C-III and D-III might be due to the requirement to release all bottlenecks at the same time.

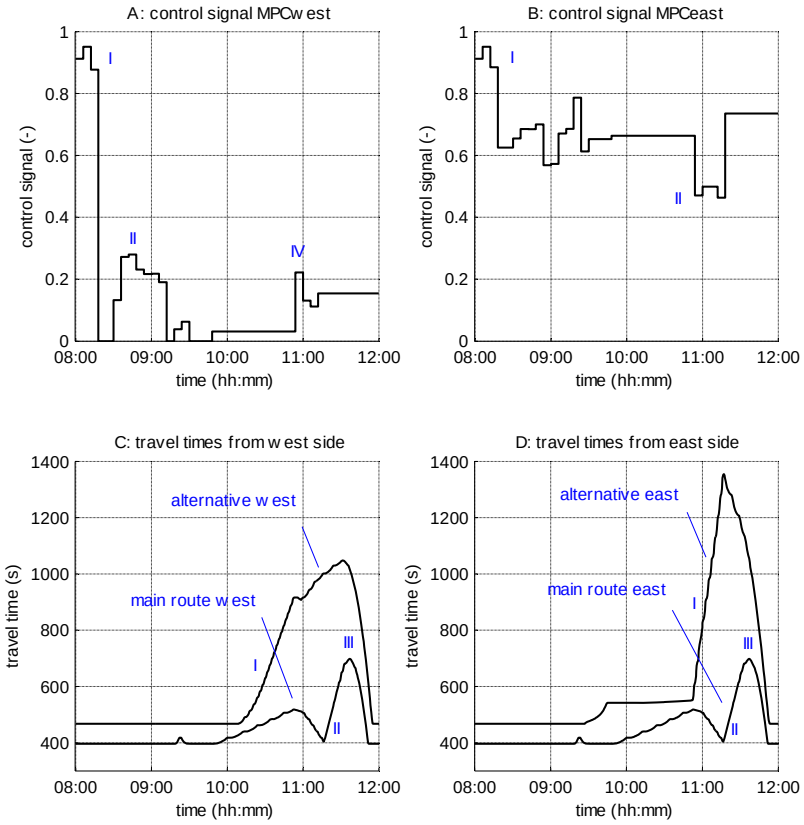

Figure 6: Control signals and travel times MPC approach

4.4 User equilibrium feedback approach

The user equilibrium approach will be evaluated briefly. The control signals to equalize the travel times over the routes are given in Figure 7A and B. In Figure 7A-I, B-I, C-I and D-I shows that traffic is first sent to the main route in order to equalize its travel time with that of the free alternatives. The signals are further chosen such that the travel times remain equal over the simulation period. The disadvantage of using this approach with respect to the network performance is that congestion is realized within the main route while redundant capacity remains available on the time-longer alternatives. Moreover, the congestion on the main route grows gradually with that on the alternatives, hence also in line with the fast travel time increase at low capacity alternatives. On important city arterials that distribute traffic over the urban area, traffic flows that do not need to pass the bottleneck can become easily hindered (like turning flow towards D_2) which has negative impact on the network performance. In Figure 7C and D this can be seen at 10:30AM where the travel times on the main routes exceed their critical value (600 seconds, see 3.4 Finite-state machine set-up) to cause hindrance to the turning flow towards D_2, while the other approaches prevent this hindrance.

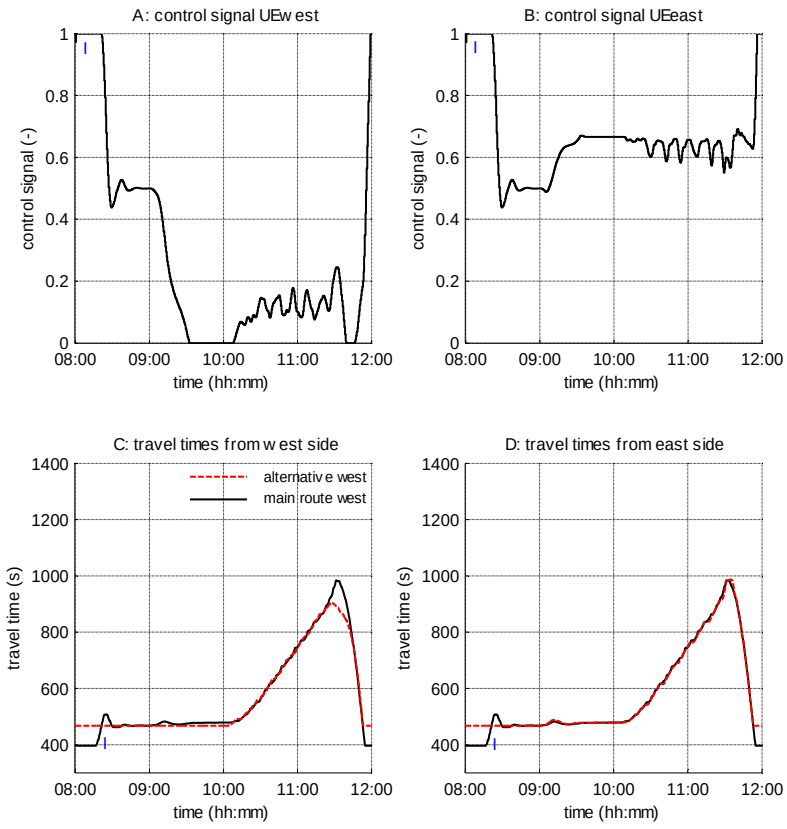

Figure 7: Control signals and travel times for the user equilibrium feedback approach

4.5 Performance and computational demand

The computational demand of the proposed approach is compared to that of the MPC based approach. However, making a comparison is difficult because both approaches differ in the way they compose the control signal and the number of times the controller is activated to realize network behavior in accordance with the objectives. An assessment is therefore made based on the computational demand for realizing the control loop once as can be seen in Figure 6A. The computational demand per control cycle of MPC is a magnitude larger than that of the finite-state machine. The computational demand of the latter is, however, practically completely determined by the travel time prediction. The jumps in the cputimes for travel time predictions indicate that some of the route travel times exceeded the initial prediction horizon, so that the horizon needed to be extended. The cputime to determine a control signal given the travel time input lies in the order of 10^{-3} to 10^{-2} seconds. This is interesting with respect to applicability and scalability of the method, since travel time predictions can be made separately from the control method. MPC on the contrary needs the predictions in its optimization procedure, meaning that its computational demand cannot simply be reduced. Moreover, the computational demand is exponentially related to the number of variables that need to be optimized. Figure 6B shows this problem when we increase the length of the control horizon from 1 to 7 per actuator.

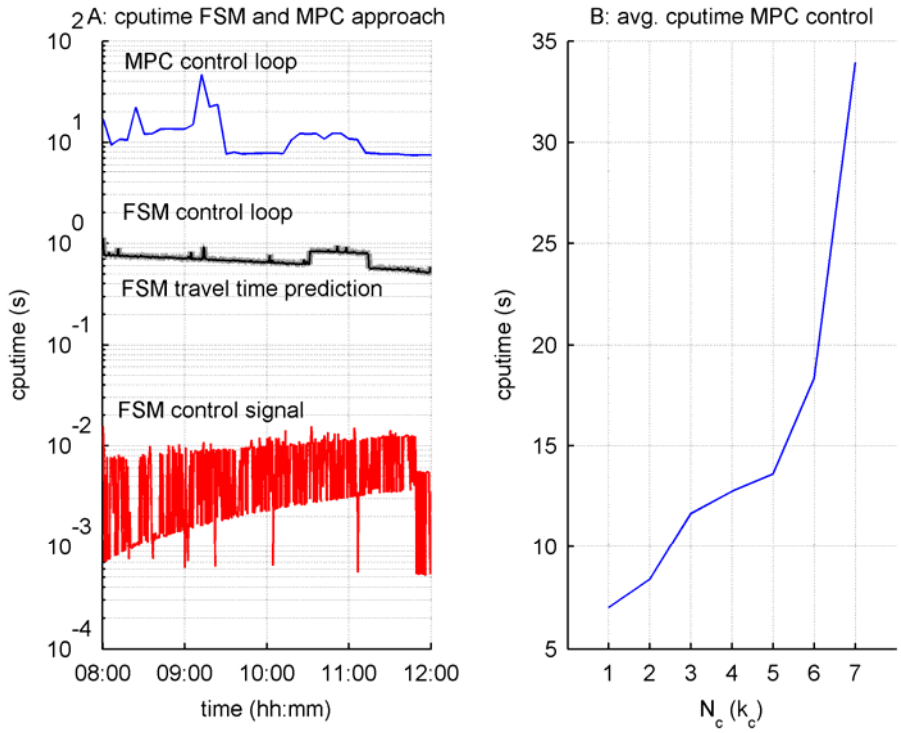

Figure 8: Computational demand of finite-state machine and MPC approach

4.6 Network performance indicators

In this section the total time spent by vehicles in the network per destination and the maximal queue lengths per route are given in Table 4 for all three approaches.

Table 4: Network performance indicators TTS and maximal queue lengths.

	TTS_{D1} (hours)	TTS_{D2} (hours)	TTS_{TOT} (hours)	$W_{\text{aWest}}^{\max}$ (m)	$W_{\text{mWest}}^{\max}$ (m)	$W_{\text{mEast}}^{\max}$ (m)	$W_{\text{aEast}}^{\max}$ (m)
FSM	1404	215	1619	2483	1000	1000	1372
UE	1440	259	1699	2000	1900	2000	739
MPC	1386	218	1604	2600	1300	1300	1267

With respect to the overall total time spent, we can conclude that the performance of the finite-state machine approximates that of the MPC based approach, and that both outperform the user equilibrium approach.

The total time spent towards D_1 is smallest for the MPC based approach because it activates and releases the bottlenecks within the routes at the same time, so that capacity is optimally used. The finite state machine is little less efficient, because it cannot adequately anticipate on this matter. Remember that the interaction mechanism required degrading one of the alternatives one service level before all available capacity became fully utilized. The user equilibrium feedback approach performs worst, because it starts to degrade the main routes till their travel times are equal to that of the corresponding alternatives before redundant capacity is utilized. During both the degradation and recovery process this leads to under utilization of available capacity.

Equalizing the travel times over all routes in that sense also causes the most hindrance to turning traffic to D_2. The finite-state machine performs best, because it prevented the turning direction to become blocked at all times, contrary to the optimal approach that allowed the turning direction to become shortly blocked (see also section 4.3). This is also indicated by the maximal queues on the main routes $W_{\text{mWest}}^{\max}$ and $W_{\text{mEast}}^{\max}$.

5 Discussion

In this contribution we have presented a service level based routes guidance approach that is able to route traffic in line with the traffic management policy objectives of the road authorities. The method degrades and recovers the routes under control stepwise, realizing system states that reflect the objectives in a systematic and comprehensible way.

The prerequisite is that the policy objectives (often qualitatively defined), the priorities of the routes and their functional requirements are carefully translated into maintainable service level bounds. The bounds can be chosen such that the controller realizes more system optimal or user optimal conditions, enabling the improvement of system performance while respecting road user interests. Network performance is improved during undersaturated conditions by utilization of redundant capacity in

route alternatives, and during oversaturated conditions by distributing the queues over the routes such that the blockage of traffic that does not move into the route's active bottleneck is prevented. User interests are respected by the definition of maximal quality differences within the service levels. Another interesting aspect of service level definitions is that it enables the realization of environmental and safety-related policy objectives, for instance, by maintaining some minimal service level in a route.

With respect to network wide implementation we can make the following remarks concerning computational demand and scalability. The computational time of MPC based approaches normally increases exponentially with respect to the number of control signals that need to be optimized, which severely limits its applicability in practice. The finite-state machine, however, does not have such high computational requirements, and the realized control signals are comprehensible and easy to interpret. Moreover, the method can be easily adopted in field of operational traffic management in The Netherlands since it is designed with respect to ongoing developments.

The test case illustrates the functioning of the control approach for a simple, yet understandable case in which the control approach is preventing blocking back from single bottleneck to upstream bifurcation points. Within more complex route layouts with multiple bottlenecks, realizing a certain service level does not necessarily mean that the blocking back phenomenon is adequately dealt with. However, degrading the performance of a high priority route in smaller steps than its alternative will at least delay the moment blocking back occurs. Moreover, the finite-state machine approach can also be used to redistribute queues over available storage space at upstream road sections. This enables a unique mapping between a travel time indication and congestion within a route. The combination of service level-oriented route guidance and this redistribution of queues within a route will then enable adequate handling of the blocking back phenomenon in a more realistic setting.

Future research includes the extension of the method to realize coordination between intersections and between a ramp meter and its upstream intersection. The finite-state machine approach then becomes the basis for service level-oriented network management and the coordination of DTM measures on a regional scale in line with policy objectives.

Acknowledgements

The research presented in this article is part of the research program "Traffic and Travel Behavior in case of Exceptional Events", sponsored by the Dutch Foundation of Scientific Research MaGW-NWO.

References

van den Berg, M., B. De Schutter, A. Hegyi, and J. Hellendoorn (2004), Model predictive control for mixed urban and freeway networks, *In Proceedings of the 83rd Annual Meeting of the Transportation Research Board*, Vol. 19.

Davis, L. (2010), Predicting travel time to limit congestion at a highway bottleneck, *Physica A: Statistical Mechanics and its Applications*, Vol. 389, No. 17, pp. 3588–3599.

Hegyi, A. (2004), *Model predictive control for integrating traffic control measures*, Ph.D. thesis, TRAIL thesis series, Delft University of Technology, Delft, The Netherlands, 2004, ISBN 90-5584-053-X.

Hoogendoorn, S. (1997), Optimal control of dynamic route information panels, *In 4th World Congress on Intelligent Transport Systems*, pp. 399–404.

Karimi, A., A. Hegyi, B. De Schutter, J. Hellendoorn, and F. Middelham (2004), Integrated model predictive control of dynamic route guidance information systems and ramp metering, *In Proceedings IEEE- ITSC*, Washington, D.C., USA, 2004, pp. 491–496.

van Kooten, J. and K. Adams (2011), *Handbook Sustainable Traffic Management Plus* (in Dutch), CROW, ISBN: 978-90-6628-576-7.

[3] Landman, R., S. Hoogendoorn, S. Hoogendoorn-Lanser, M. Westerman, and J. Van Kooten (2010), Design and Implementation of Integrated Network Management in the Netherlands, *In Proceedings of the 89th Annual Meeting of the Transportation Research Board*.

Landman, R., A. Hegyi and S. Hoogendoorn (2011), Service Level-Oriented Route Guidance in Road Traffic Networks, *In Proceedings of IEEE-ITSC*, pp. 1120-1125.

Landman, R., T. Schreiter, A. Hegyi, van Lint, J. and S. Hoogendoorn (2012), Policy-based service level-oriented route guidance in road networks: a comparison with system and user optimal route guidance, *In Proceedings of the 91th Annual Meeting of the Transportation Research Board*.

van Lint, J., S. Hoogendoorn, and M. Schreuder (2008), Fastlane: New Multiclass First-Order Traffic Flow Model, *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2088, pp. 177–187.

van Lint., J. (2010), Empirical Evaluation of New Robust Travel Time Estimation Algorithms, *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2160, pp. 50–59.

Mammar, S., A. Messmer, P. Jensen, M. Papageorgiou, H. Haj-Salem, and L. Jensen (1996), Automatic control of variable message signs in Aalborg, *Transportation Research Part C: Emerging Technologies*, Vol. 4, No. 3, pp. 131–150.

Messmer, A. and M. Papageorgiou (1995), Route diversion control in motorway networks via nonlinear optimization, *IEEE Transactions on control systems technology*, Vol. 3, No. 1, pp. 144–154.

Messmer, A., M. Papageorgiou, and N. Mackenzie (1998), Automatic control of variable message signs in the interurban Scottish highway network, *Transportation Research Part C: Emerging Technologies*, Vol. 6, No. 3, pp. 173–187.

Minciardi, R. and F. Gaetani (2001), A decentralized optimal control scheme for route guidance in urban road networks, *Proceedings of IEEE-ITS*, pp. 1195–1199.

Papageorgiou, M. (1990), Dynamic modelling, assignment, and route guidance in traffic networks, *Transportation Research Part B: Methodological*, Vol. 24, No. 6, pp. 471 – 495.

Pavlis, Y. and M. Papageorgiou (1999), Simple decentralized feedback strategies for route guidance in traffic networks, *Transportation Science*, Vol. 33, No. 3, pp. 264–278.

Peeta, S. and H. Mahmassani (1995), System optimal and user equilibrium time-dependent traffic assignment in congested networks, *Annals of Operations Research*, Vol. 60, No. 1, pp. 81–113.

Rijkswaterstaat (2003), *Handbook Sustainable Traffic Management*, AVV Transport Research Centre, ISBN 90-3693-625-X.

Wahle, J., A. Bazzan, F. Klügl, and M. Schreckenberg (2000), Decision dynamics in a traffic scenario, *Physica A: Statistical Mechanics and its Applications*, Vol. 287, No. 3, pp. 669–681.

Wang, Y., M. Papageorgiou, and A. Messmer (2003), Predictive feedback routing control strategy for freeway network traffic, *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1856, pp. 62–73.

Wang, Y., M. Papageorgiou, and A. Messmer (2001), Feedback and iterative routing strategies for freeway networks, *In Proceedings of IEEE-CCA*, pp. 1162–1167.

Zuurbier, F., H. Van Zuylen, S. Hoogendoorn, and Y. Chen (2006), Generating Optimal Controlled Prescriptive Route Guidance in Realistic Traffic Networks; A Generic Approach, *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1944, pp. 58–66.

Zuurbier, F. (2010), *Intelligent Route Guidance*, Ph.D. thesis, TRAIL thesis series, Delft University of Technology, Delft, The Netherlands, ISBN 978-90-5584-131-X.