Operational Control Solutions for Traffic Management on a Network Level

Landman, R.L.

DOI
10.4233/uuid:02596b75-8108-444f-957d-dff4bf9226fa

Publication date
2021

Document Version
Final published version

Citation (APA)
Landman, R. L. (2021). *Operational Control Solutions for Traffic Management on a Network Level*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:02596b75-8108-444f-957d-dff4bf9226fa

Operational Control Solutions for Traffic Management on a Network Level

Ramon Leonardus Landman

The research leading to this dissertation has received a VICI funding (Innovational Research Incentives Scheme) from NWO - the Dutch Organisation for Scientific Research in the Councel area MaGW for Social and behavioral sciences in 2008.

Operational Control Solutions for Traffic Management on a Network Level

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, Prof.dr.ir. T.H.J.J. van der Hagen
chair of the Board for Doctorates
to be defended publicly on
Monday 14, June 2021 at 10:00 o'clock

by

Ramon Leonardus LANDMAN
Master of Science in Civil Engineering
Delft University of Technology, the Netherlands
born in Rotterdam, the Netherlands

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof.dr.ir. S.P. Hoogendoorn	Delft University of Technology, promotor
Dr.ir. A. Hegyi	Delft University of Technology, copromotor

Independent members:

Prof.dr.ir. J.W.C. van Lint	Delft University of Technology
Prof.dr.ir. E.C. van Berkum	University of Twente
Prof.dr. L. Leclercq	Gustave Eiffel University, France
Dr. F. Viti	University of Luxembourg, Luxembourg
Dr. M. Menendez	NYU Abu Dhabi, United Arab Emirates

TRAIL Thesis Series no. T2021/17, the Netherlands Research School TRAIL

TRAIL
P.O. BOX 5017
2600 GA Delft
The Netherlands
E-mail: info@rsTRAIL.nl

ISBN: 978-90-5584-292-6

Printed in the Netherlands

There is also such thing as a spirit of the times, an attitude of mind characteristic of a particular generation, which is passed on from individual to individual and gives a society its particular tone. Each of us has to do his little bit towards transforming this spirit of the times.

Albert Einstein

Acknowledgement

As I look back on a decade of life lessons brought to me by science and due to science, many faces cross my memory. What a ride! It all started in 2009 with the invitation of becoming part of a wildly ambitious initiative to test the potential of integrated network management in practice - a learning-by-doing type of approach embedded in the Field Operational Test Amsterdam. As I like learning new things, that sounded like fun. However, I had no idea what I was getting into and how the whole experience of becoming a doctoral candidate would turn my world and my world view upside down forever.

The Amsterdam project tested state-of-the-art traffic control and monitoring techniques and their applicability in practise. In parallel efforts were made to keep stakeholders involved at all times and I got extremely interested into the organisational hassle of it all. My love for technical challenges and people was born. I became amazed by the fact that people from different sectors can have so much trouble truly understanding each other and that even with a lot of expertise sitting at the table issues were not necessarily solved. It was shockingly revealing that there is a whole different level to traffic network management called politics. Either way, with so many layers to discover and explore it didn't take long to become neck-deep involved.

Being offered the position at Delft University of Technology as a researcher and scientist was truly life changing, so the first big acknowledgement goes to Serge Hoogendoorn. The man who has thought me how to stay on track and stay focussed, who has showed that very complex problems can be solved by elegant and genius solutions, and how a short memo can be used to start or push forward a discussion and create more support. I am deeply grateful for you preparing me to apply science in a more playful way as much suited to reality as possible and needed. I would like to thank you for sharing your insights, the many talks we have had, the travels we

have made and the many laughs we shared in between the bumpy parts of my journey. Hanging around with you has given me a beautiful imprint on how to build bridges between practice and science. Apart from getting to know myself and my limits, whilst putting science into practice in a pressure cooking environment, I learned really cool stuff. And there's one person to thank for especially: Andreas Hegyi. He taught me how to build algorithms based on conceptual thoughts, how to get to the essence and develop and test innovative concepts by means of different types of simulation environments. Your guidance and scientific wisdom pushed me to the very end of my ability and led me to develop a more curious attitude with an eye for detail and a more fundamental way of doing research. The mantra involved back then, I still live by today: start simple with a minimum not trivial case and add more complexity along the way. In that way you can always find your way back to conceptual flaws. In addition I would also like to thank the other members of my committee for their quick responses and heartfelt reactions: Prof. E.C. van Berkum, Prof. L. Leclercq, Dr. M. Menendez, Dr. F. Viti and Prof. J.W.C. van Lint.

So here we are in 2021, after all these years wrapping a ribbon around it all. There are so many people to be grateful for. People who in their own way held space for me and therefore kept the research alive or that have given support in reaching its closure. There are a few I want to personally address. First up: Miriam. Thank you for keeping our household going back then; for attending our daughters when I was caught up in work or even abroad. Thanks for being there while I was (at times desperately) trying to strip a problem to its essence and identify its root cause. And by doing so, thank you for giving me the opportunity to invest in myself. Then, my dearly beloved parents: In all these years you never stopped asking me when I would finish and at the same time you have always made it clear it would have been okay if I did not. This heart-warming attitude is what enabled me to stay open and not shut the door to finishing this doctorate. And, Sis, I also thank you for your loving attention. Being a guinea pig within your NLP, mBIT, and Shiatsu courses has been very encouraging throughout my journey. Another family member that has always been there over the years is my nephew Jelle. After all these years and experiences, I can conclude that Zeeland became my second home. I love to lose my mind and come back to my senses in your grand hotel. Mate, thanks for all the good laughs, fun and

your care - seeing the sun go down with a bottle of wine at our bench on the beach will always remain magical.

I would also like to express my gratitude for the ever-welcoming Maaike & Gerard. Our talks, dinners, the Combe cards and your ever sincere interest in my being and doing have supported me in more ways than you know. Having really good friends is a most definite need when tripping towards a finish line. In retrospect there have been quite some people showing genuine interest in my research which sparked (at times my more hidden ambition) to really draw my conclusions and finish this book. I will name one in person: Remco, thank you too for giving me the encouragement to keep my research alive.

My last acknowledgement I dedicate to the love of my life: Nina, with whom I share a deep-rooted joy for bringing the essence of things to light in a holistic way. On one of our first dates, she told me she had dug up my scientific articles and read them. This I found shocking, but then we enrolled in a long talk about my research: why it was done; what it could serve for, and what would be needed to finish it. One week later, on a Sunday morning she interviewed me about my own conclusions and recommendations and that was the start of respectfully closing this research. I want to thank you for underlining the importance of presenting holistic views on matters that are important to public policy. It helped me to truly believe in the relevance of putting this research into the bigger perspective. My love, you held my hand, at times kicked my ass and walked me to the finish line - thank you.

As you can read, I dived deep and surfaced again and again. Over the years I became a more independent thinker and I like to believe this all helpt building more stamina when being outside my comfort zone trying to make sense of things and looking for root causes. Last but not least, I would like to express my wish that when my daughters Inoa and Elyse are a certain age, they will read this acknowledgement, maybe even the whole book and see, feel and learn that there is no greater good than to keep personally growing with enough guidance and support around you.

Contents

Chapter 1

Introduction

1.1 Motivation

Roads are indispensable for people to perform both economic and social activities. However, due to the ever growing traffic demand, which is generated by the mobility of people and the transportation of goods, urban areas all over the world are facing serious congestion problems. When the traffic demand exceeds a network's supply, phenomena are triggered that decrease the network capacity. The irony being, that the capacity is most needed at these specific moments. Examples of such phenomena are the freeway capacity drop when congestion sets in, blocking back of queues over upstream bifurcation points and suboptimal route choice of vehicles within the network. Apart from the fact that the resulting congestion causes considerable costs due to unproductive time loss, it also increases incident probabilities, pollution, and the impact on the environment and the quality of life.

The current focus with respect to mitigating these negative impacts, is the application of dynamic traffic management measures on a network level. Congestion problems have for a long time been solved by extending the road infrastructure. However, building new roads is a very costly solution that has a large impact on the living environment. The focus has therefore been put on the development of dynamic traffic management measures that enable more efficient and safe use of the existing network infrastructure. Local traffic management measures were developed to postpone or prevent the onset of the earlier mentioned traffic phenomena, such as ramp metering to pre-

vent freeway flow breakdowns, signal control to maximize the throughput at intersections and route guidance to use all available network capacity. However, nowadays demands are frequently exceeding the capacity at more and more locations within the network. This implies that by solving one bottleneck, others might be easily activated elsewhere in the network. Besides, with multiple active bottlenecks, control measures need to address them in such a way that the network performance is maximized. The assessment of the spacial relation between bottlenecks and the impact of control measures has therefore become very important.

Essentially, the onset of undesired traffic phenomena can be postponed by either temporarily holding back traffic that moves towards the bottleneck at strategic locations elsewhere in the network (e.g. at storage spaces such as controlled on-ramps and arms of an intersection), or by rerouting traffic over the network parts where redundant capacity is available. The duration of which traffic problems can be prevented is in this respect directly related to the available amount of storage space or capacity in alternative routes. In other words, the more space or alternative route capacity is available, the longer the inflow to a bottleneck can be reduced. To hold back traffic elsewhere in the network, traffic control measures usually need to cooperate; this means realizing *coordination* between measures of the same type and *integration* between measures of a different type. Aspects in this respect, that affect the network performance are the allocation of coordination storage spaces and the strategy with which they are used.

Designing and operationalizing good control strategies that maximize the network performance is not easy. Optimal control strategies were initially explored by means of approaches that use a model and optimization procedure to minimize objectives such as the total time spent by vehicles in the network. However, large-scale applicability of these approaches is limited due to the following reasons. Their computational demand is high and control signals can become suboptimal or even counter productive in case there is a mismatch between the prediction model and reality. Moreover, evaluating the controller behavior and its performance is difficult, if not impossible, when optimizing the signals of many actuators in a complex and large-scale network setting with unknown demands. Therefore, the knowledge that is gained on optimal strategies by means of minimal, but not trivial simulation

test cases, is still predominantly used for the development of heuristic approaches that realize control objectives in a more comprehensible way.

In the operational field of traffic management there are more objectives to satisfy than just improving the overall network performance. This is related to the many different stakeholders involved in the process of formulating a vision upon the functional use of the network. From the point of view of the network operator and the network users typical objectives are:

- **Improving the network performance.** By preventing phenomena that decrease the network outflow, vehicles will spend less time to complete all trips. However, despite the fact that the *average road user* will benefit from increased network performance, preventing bottlenecks by means of control might have a serious impact on the travel time of the individuals that are delayed.

- **Maintaining a certain level of equity.** It is therefore important in an operational environment, that travel time loss encountered by the individual road users remains acceptable. This can be influenced by the users themselves using navigation systems, or by the authorities that set limitations on allocated storage spaces and maximum travel time differences over controlled route alternatives.

- **Limiting the traffic impact on safety and environment.** Environmental and safety objectives can be satisfied by means of measures such as lowering the speed of traffic flows or by preventing congestion. However, reducing the speed of vehicles will also result in reduced network performance or outflow. This implies that political decisions need to be made when interests result in conflicting desires on the network utilization.

The interests of stakeholders need to be harmonized to realize a common vision upon the functional use of the network that can also be operationalized. This includes making agreements on desired performance or service levels, that reflect objectives with respect to equity, safety and the environment, given aspects such as the roads' type, the time of day and typical demand characteristics. Aiming at improving the network performance while respecting policy objectives that pose restrictions on achieving system optimality, can therefore be considered a realistic format for operational

network traffic management. With the stakeholders agreeing on a common vision, systems are needed that are able to operationalize the vision based on real-time conditions at the involved freeways and urban roads.

1.2 Research objective and questions

Driven by the aforementioned issues and requirements, this thesis aims at designing new traffic control strategies and algorithms that improve the network performance, while equity, safety and the environmental related objectives of the road authorities are satisfied. To come up with such system, the following research questions are studied:

- What traffic phenomena affect the performance of a road network?
- How can the impact of these phenomena be limited?
- How can the most recent control strategies and methods be improved?
- How can traffic control strategies operationalize road authorities' policy objectives?
- How must buffers be selected and deployed to maximize the improvement of the network performance?
  - Where do the costs and benefits occur in terms of travel time?
  - How can we gain insight into these costs and gains to improve our control strategies?
- What preconditions does operational practice impose on coordinated network-wide traffic management?
  - How can comprehensibility of the control framework and its actions be assured?
  - How can complexity and computational demand remain limited for real-time operationalization?
  - How can unforeseen changes in traffic demand and infrastructure be dealt with?
  - How can we systematically setup and tune the parameters of the system?

- How can generic methods for data collection and traffic management be integrated at network level?

- What are the preconditions for designing an integrated traffic management system?

- What are the lessons learned in respect to the design and implementation of the large-scale field operational test for traffic management in Amsterdam?

1.3 Research approach

In order to define the research needs more clearly, we start with a literature survey on existing control approaches that are able to redistribute traffic such that the network performance is improved. The focus is put on the deployability of the approaches and their ability to realize target states in line with policy objectives such as network performance and equity.

Based on the findings from literature, preferred controller types (i.e. feedback, feedforward, optimal control) are selected that are suitable for application in an operational context. Then, existing control approaches will be evaluated to see if there are opportunities for further improvements with respect to more efficient network utilization in line with policy objectives.

The focus is put on the development of effective heuristic control strategies that are based on a thorough understanding of the root causes of and the solutions for undesired traffic phenomena in the network. Although optimal control-based approaches are superior in terms of flexibility and robustness, for practical applications the computational complexities and lack of transparency are sufficiently serious disadvantages to consider heuristic but still generic approaches. Optimal control can nevertheless be used to benchmark the performance of the proposed heuristic approaches.

Since traffic networks consist of both freeway and urban roads, coordinated or integrated control solutions will be designed for both types of networks. The coordinated and integrated control solutions need to be efficient,

comprehensible and easily superimposed on existing control measures such as ramp meters and intersection controllers.

Dynamic control and monitoring units have many parameters to be set in order to adequately react on traffic conditions. Where needed, approaches will be considered or designed to systematically tune and configure the system parameters. This limits the time and effort needed for setting up the system and it will result in improved system performance.

The aim is to integrate the control and monitoring solutions in an overall network management system that can be tailored to the specific needs of any regional network. It is important that new monitoring and control units can be easily adopted and that the system can be stage-wise operationalized in practice.

Finally, designing an integrated network management system includes the use of traffic simulation models for the purpose of understanding, testing, improving and validating the overall control system and its individual components. In order to ensure their well-functioning a structured and feedback-based design flow will be defined that keeps the conceptual, technical and functional tests on the system comprehensible and manageable.

1.4 Research scope

Within this thesis we will focus on defining different integrated and coordinated control strategies based on some of the most widely applied traffic control measures: route guidance, ramp metering and intersection control. Despite the fact that measures such as dynamic speed limits and in-car devices would also fit well within such control framework, they remain out of the scope of this research.

The potential effectiveness of the proposed control approaches are discussed based on simulation experiments. The test cases mainly consist of minimal -but not trivial- network layouts and simple demand patterns that enable unambiguous hypothesis testing on desired control behavior in the context of typical scenarios. The used simulation environments are mostly macroscopic and deterministic in order to easily reproduce outcomes. This

also means that the outcomes are not biased by modelling artefacts and stochasticity that can not be reproduced and that make it difficult to interpret what the control approach is doing.

The proposed strategies consider vehicular traffic for the urban and freeway networks only. Hence, road users such as pedestrians and cyclists are not considered in the design of the control strategies.

No claims are posed with respect to the potential of network traffic management in general, for this is highly dependent on aspects such as network and demand characteristics, the type of controllers, how well these are tuned, and policy objectives that constrain achieving system optimality.

1.5 Theoretical contributions

The theoretical contributions of this thesis consist of operational control approaches and different analytical evaluation approaches for:

- **Service level-oriented route guidance.** A dynamic feedback-based route guidance approach is proposed to operationalize policy objectives that are reflected by target service level values. During oversaturated conditions where there is more traffic demand than route capacity, the routes' service levels are degraded by means of a finite-state machine, such that the decreased network performance is postponed while the travel time difference over the routes remains within certain equitable boundaries. By combining classic control theories such as rule-based switching schemes (i.e. the finite-state machine) with state feedback control laws, more complex traffic control strategies can be deployed that involve the execution of multiple control tasks at the same time. This means that desired target values for stable network states can be maintained at multiple locations given prevailing conditions.

- **Coordinated ramp metering.** A ramp metering strategy has been developed that enables the utilization of ramp storage space such, that a freeway flow breakdown and the associated capacity drop are maximally prevented. To this aim, more upstream-located ramps are always saturated before more downstream-located ones, i.e. the ramps

run out of space in downstream direction to make sure that all realized assistance is used by the ramp that is metering on the bottleneck. Moreover, an operational algorithm has been designed that is able to actualize this strategy in real-time, based on the state estimations of the involved freeway and its connecting on-ramps.

- **Coordinated intersection control.** To enable longer ramp metering control on a freeway bottleneck or to prevent blocking back of queues within an urban arterial, a coordinated intersection control strategy has been developed that temporarily stores vehicles at upstream-located intersections from the bottleneck. Based on simple state feedback control laws, all buffers at the coordinated intersections are filled simultaneously (i.e. in parallel).

- **Integrated ramp metering and intersection control.** A coordinated control approach is proposed to fill storage spaces at the intersections upstream of a ramp one after the other (i.e. sequentially) to longer prevent ramp saturation while minimizing waiting delays. As for the service level oriented route guidance approach, this approach also makes use of the combination of a rule-based switching scheme (i.e. the finite-state machine) and state feedback control laws that enable controlling the state at multiple buffers simultaneously.

- **Storage space allocation and utilization.** An evaluation approach has been developed to a priori identify the optimal set of coordinated storage spaces when applying coordinated ramp metering or integrated control between a ramp and its upstream located intersections. To this aim, cumulative curves are defined as a function of the network, demand and control characteristics for the key locations in the network where delays occur. The curves enable us to search for the storage space configuration that minimizes the overall system delay. Moreover, insight is gained into the variables and key mechanisms that determine the benefits and costs of coordinated and integrated control approaches when postponing freeway flow breakdown. The approach can also be used to determine the optimal strategy with which the coordinated storage locations are filled (e.g. sequentially or in parallel).

- **Systematic tuning approach for feedback gains.** A systematic tuning approach is presented to analyze the stability behavior of state feedback control laws. The approach enables finding optimal settings

for the involved feedback gains of the feedback controllers. By writing the system dynamics including the control laws in state space form, an eigenvalue analysis can be performed that identifies which feedback gains result in either a stable or unstable system.

1.6 Practical contributions

The practical contributions of this thesis are:

- **Network management system architecture.** A generic and modular framework is proposed for the integrated operation of traffic management measures within a regional network. New approaches can be easily adopted that improve the effectiveness of the overall control system. Its generic and modular setup also allows road managers to tailor the system to specific network layouts, traffic problems and policy objectives.

- **Structured design process.** Guidelines are presented for the use of traffic flow models when designing and testing new control and monitoring strategies and their integration within a control framework. These guidelines can be used by technicians and project managers to better manage the design workflow of such system and to waste as little time as possible during the conceptual, technical and functional testing phases.

- **Valuable lessons learned.** To preserve the knowledge for future realizations of network management, the lessons learned are elaborated on the process from designing the control and monitoring units, to their integration in the control architecture and their operationalization in practice.

1.7 Thesis outline

The content of each chapter and how they relate are schematically outlined in Figure 1.1. The starting point of the work presented in this thesis is the situation where there is a common vision available on how involved stakeholders want their regional road network to function. Subsequently, control

systems are needed to operationalize this vision based on real-time conditions at involved freeway and urban networks.

First, in **Chapter 2** an overview and discussion is presented of previously proposed local, integrated and coordinated control approaches with respect to route guidance, ramp metering and intersection control. The review identifies our problem statement for further research and important design aspects that need to be taken into account when developing an operational network management system.

Given our starting point, the most obvious way to redistribute traffic over the network would be by means of route guidance in line with the policy objectives of the road authorities. In **Chapter 3** a route guidance approach is discussed that degrades the service levels of different routes between an origin and destination pair stepwise in line with the predefined target values and given the prevailing traffic conditions.

In case the route guidance approach is used to maintain certain service levels in terms of travel time, it is impossible to adequately prevent traffic problems when there are multiple locations within a route where delays are encountered. Travel time does not give an unambiguous state estimate that identifies location specific bottleneck. To nevertheless prevent location specific phenomena such as the capacity drop and spill-back within a route, *traffic that moves towards the bottleneck* can be temporarily stored in the vicinity of the bottleneck. To this aim, coordination between traffic management measures needs to be realized.

In **Chapter 4** it is discussed how to allocate these storage spaces such that the total system delay is minimized when temporarily holding back traffic to postpone undesired traffic phenomena. Moreover, we also elaborate on different storage space utilization strategies and their impact on generated delay.

Subsequently, coordination algorithms are needed to enable the use of these allocated storage spaces in line with the above strategies. A coordinated ramp metering approach is described in **Chapter 5** to use coordinated upstream ramp storage space to prevent freeway flow breakdowns. The al-

gorithm saturates the assisting ramps in downstream order to increase the efficiency with which their storage spaces are used.

In **Chapter 6** we present different algorithms for using the storage space in urban arterials to prevent spill-back of ramp queues and intersection queues during oversaturated conditions. More specifically, the first algorithm fills allocated storages spaces at intersection controllers in the urban arterial in *parallel*, meaning they are all filled simultaneously. The second algorithm is able to also *sequentially* fill the storage spaces to minimize the hindrance to stored vehicles that do not need to pass the bottleneck. The algorithms are of the state feedback type, hence in this chapter we also discuss a method for systematically tuning the feedback gains.

In **Chapter 7** the control architecture is discussed that was designed for the field operational test in Amsterdam to realize network management in practice. The corresponding control and monitoring units will be shortly introduced to illustrate that the framework allows for the addition of any type of control and monitoring unit.

When designing the different control and monitoring units, models are used to evaluate their design and to test their technical implementation and functional behavior. **Chapter 8** elaborates on how such process can be effectively shaped and on some valuable lessons learned.

Finally, **Chapter 9** summarizes the main conclusions with respect to answering the research questions and achieving the objective of this thesis. Moreover, the implications for the stakeholders and the recommendations for future work on this topic are presented.

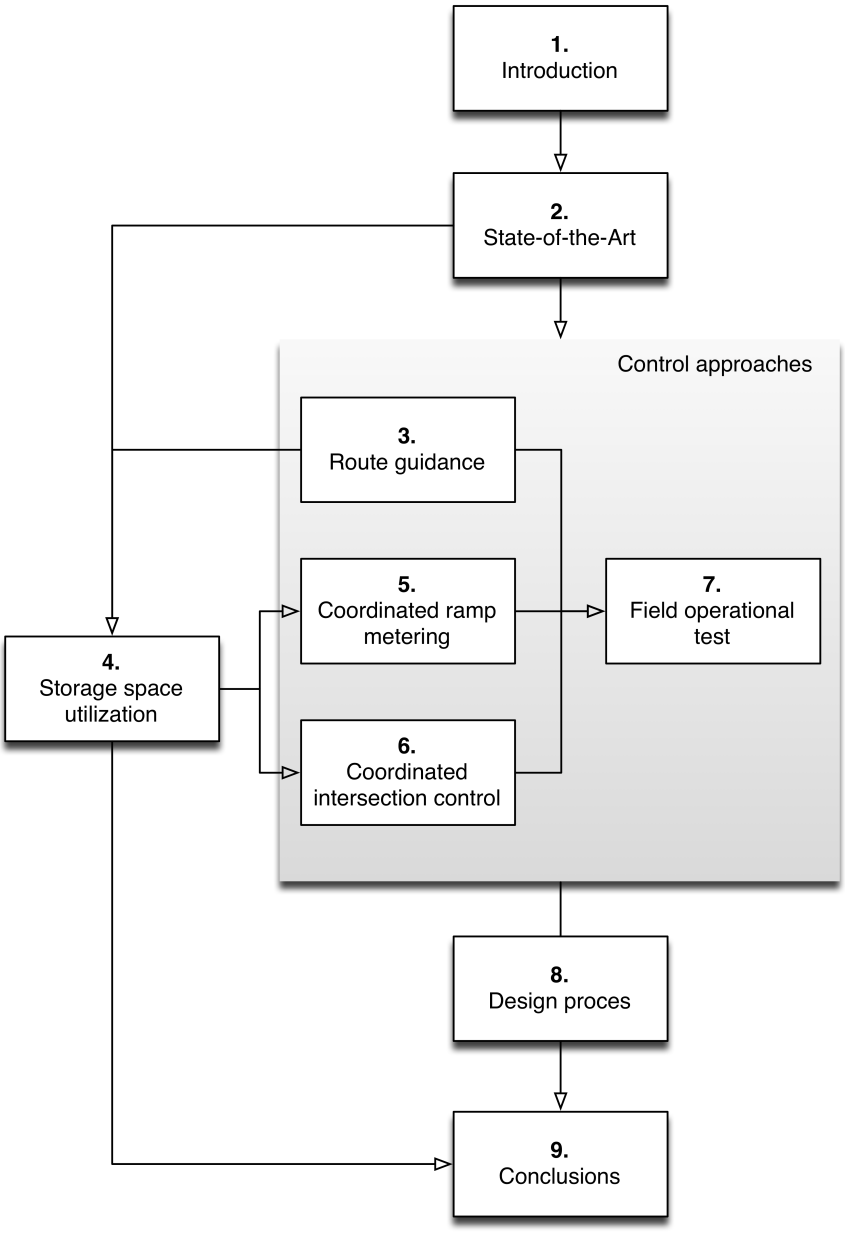

Figure 1.1: Thesis outline.

Chapter 2

Background on local and coordinated traffic control

In this chapter a literature and state-of-practice overview is given about the technical and organizational developments with respect to solving traffic problems on a network level. The focus will be put on the three most well-known and widely applied forms of traffic management, being route guidance, ramp metering and intersection control. Moreover, apart from their local application, different forms for coordination and integration between these measures are addressed to identify the niches for further research.

2.1 Introduction

In this literature and state-of-practice overview the transition is discussed from solving traffic problems locally to solving them from a network perspective. Since the 1970s, control strategies have been under development that enable traffic management measures to collaborate on a network level. By realizing coordination between traffic management measures, traffic is distributed over the network such, that the onset of phenomena that decrease the network performance is postponed. This field of research is still very active, which has probably to do with the ever growing congestion problem and the fact that the transition to network traffic management involves challenging technical developments, as well as organizational challenges. In the remainder both technical and organizational developments will be briefly addressed.

The technological focus within this overview will be on the three most well-known and widely applied forms of traffic management, being route guidance, ramp metering and intersection control. Different forms for coordination and integration between these measures will be addressed to identify the niches for further research. The most important programs, projects and applications involving integrated and coordinated control of ramp metering and intersection control are projected on a timescale to gain better insight into the developments over time. Despite the fact that dynamic speed limits and in-car devices also fit well within such control frameworks given their abilities to improve the network performance Hegyi et al. (2009); Hegyi & Hoogendoorn (2010); Mahajan et al. (2015); van de Weg et al. (2014), they remain out of the scope of this research and are not discussed in this background overview.

First, in Section 2.2 the international initiatives are identified that introduced coordinated and integrated control concepts in practice. Most of them involved large-scale field operational tests as demonstration cases around Europe and the United States. After the first successes, governments started to write policy for the realization of network management in their urban areas. In this respect, the main developments within the Netherlands are discussed in Section 2.3 to illustrate organizational changes that are needed to structurally facilitate network management on a national scale. In Section

2.4 some interesting aspects of the technical and organizational developments are addressed. To conclude, the literature on route guidance, ramp metering and intersection control are discussed in Sections 2.5, 2.6 and 2.7 to identify the research needs and technical requirements when dealing with the design of control strategies and approaches.

2.2 Transition to network wide traffic control

For a long time it is known that the economic well-being of modern societies is dependent on reliable and efficient operation of its physical socio-economic large-scale systems, such as transport systems, power networks and industrial plants. Improving the operation, reliability and productivity of such systems, will have a significant economic impact on the society as a whole Athans (1978). Despite this awareness, developments with respect to integrated and coordinated control strategies to improve the road network performance initially were very theoretical and took place outside the operational field van Aerde & Yagar (1988). This was caused, on the one hand, by the lack of simulation environments that could properly model freeways and urban traffic corridors to gain better understanding of these new control strategies and to explore their potential and operational design. On the other hand, conflicting organizational interests made operationalization in practice difficult, i.e. the fact that the urban and freeway networks are governed by different authorities with conflicting interests Taale & Westerman (2005); Mac Carley et al. (2002); Urbanik et al. (2006). Over time, multiple simulation models and optimization based control approaches were proposed to explore the potential benefits of coordinated and integrated control Papageorgiou (1995); Stephanedes & Chang (1993); Taale et al. (1994). This facilitated the leap towards field operational tests in which concepts could be actually tested.

The European Union stimulated further development of coordinated ramp metering by means of the framework programs in projects such as CHRISTIANE ('87-'91), EUROCOR ('92-'94) and DACCORD ('96-'99) Middelham et al. (1995); Kotsialos et al. (1998). In the same period, researchers in the United States worked on the design and evaluation of an integrated corridor-level adaptive control system Mac Carley et al. (2002). These programs all indicated that it is difficult to operationalize integrated and coordi-

nated control approaches, but despite many set backs some first operational successes were achieved in the late 90s. From that moment on the European research continued to stimulate the developments in projects including large-scale field operational tests around Europe as TABASCO ('96-'98) on integrated control including route guidance, SMART NETS ('01-'04) on traffic responsive urban control, RHYTHM ('01-'04) on local traffic responsive ramp metering, EURAMP ('02-'06) on local and coordinated ramp metering, and NEARCTIS ('08-'13) on creating an European network of excellence. In 2005 an ITS program was launched in the United States to develop a uniform format for operationalizing integrated corridor management Gonzalez et al. (2012). Contrary to the focus on the development of real-time control applications in Europe, the Unites States focussed more on the development of near real-time multimodal decision support systems for integrated corridor management.

2.3 Developments in the Netherlands

In the remainder of this section the consecutive developments within the Netherlands will be reviewed. The main reason being, that most of the work performed for this thesis is done in light of these developments. After a successful demonstration of coordinated ramp metering along the A10 west around Amsterdam within the European framework projects in the late 90s Kotsialos et al. (1998) and the build of a Central Traffic Management System that could communicate with the actuators in the field, the transition started to ready the organizational aspects around network management in the Netherlands.

In 2003 a method called 'Sustainable Traffic Management' was published that harmonizes the different interests of involved stakeholders Rijkswaterstaat (2002). The output of the method is a common vision on the network functioning and all of its elements, expressed in terms of road priorities and corresponding desired service levels. In 2004 a simulation model was made available to evaluate the impact of introducing network management in the larger urban regions Taale & Westerman (2005). An additional handbook would appear in 2011, on how these plans could be operationalized by means of existing traffic management measures that are available in a region van Kooten & Adams (2011). However, before the publication

of this handbook, the Dutch government would first explore the potential of network management by means of different field operational tests.

To this aim, the program FileProof ('06-'09) was initiated, to finance initiatives that would have a positive impact on the road network performance in the short term. In the project 'Verbetering Doorstroming A10' (VDA10 - in English: Improving Throughput A10) a system was developed to control the freeway and urban road network around Amsterdam in line with the formulated policy objectives by means of automatically generated control scenarios Wang et al. (2009, 2010). Based on prevailing conditions, the system would autonomously suggest the activation of certain control measures, but the operator in the traffic center would need to approve them before execution. This system came as an answer to the situation where the road operator had to select and employ scenarios manually for the complete network Volp et al. (2006); the management and maintenance of more then 700 scenarios turned out to be a rather complex and time consuming task. The extension of the traffic management arsenal in the Amsterdam region with up to 32 ramp meters, integrated with in total 57 upstream located intersection controllers, many additional variable message signs and several infrastructural adaptions turned out to reduce the delays at the freeway with more than 10% van der Veen et al. (2010). However, the field results of the automated control scenario's and their impact on the freeway and urban network performance have never been published.

Parallel to this project, a proof of concept was developed in 2009 to realize a fully autonomous integrated network management system van Kooten & Meurs (2009). The resulting concept would become the basis of the system that has been designed in the 'Praktijkproef Amsterdam' (PPA - in English: the Field Operational Test Amsterdam). To formulate adequate policy around these promising developments in 2011, the research department 'Kennisinstituut Mobiliteit' (KIM - in English: Knowledge institute for mobility policies) of the ministry studied the potential impact of different measures that enable more efficient use of the Dutch roads Savelberg & Korteweg (2011). This also lays the foundation for further investments in the development of network wide traffic management by the Dutch government.

A strategic board, 'Strategisch Beraad Verkeer en Vervoer' (SBVV - in English: Strategic board traffic and transport) was put into place SBVV (2011), consisting of influential people from the authorities and industrial parties to advise the government on its future course with respect to traffic management and information. Different future scenario's were built and evaluated, resulting in a transition where public and private stakeholders work on integrating means for informing, guiding and steering traffic.

- A scenario in which private companies solely inform road users and authorities merely set boundaries, will probably result in (severe) underutilization of the network capacity, i.e. not having a guiding road operator makes the system vulnerable to disruptions;
- A scenario in which the road operator fully controls the use of the road network (e.g. slot allocation, routing, platooning) is also not desired. Despite the fact that the reliability of the system would be increased and societal goals more easily achieved, this scenario is deemed unrealistic in light of current individualized society and the corresponding technical demands for realization.

Based on the desired future situation, the following developments were advised by the board:

- Stimulate dialog between public and private stakeholders; the development of a long term vision will stimulate investments in innovations and based on good understanding of available knowhow policy can be improved;
- Arrange a national programmatic approach for projects to embed previously gained knowledge in projects and field operational tests into new developments;
- Solve knowledge gaps on data use, e.g. juridical questions, intellectual property, liability and privacy;
- Harmonize with European developments to prevent standardization and legislation issues, and to strengthen the economical position.

The aims of making better utilization of the existing road infrastructure continued to reflect in the program Beter benutten ('11-'14). This program financed the design and implementation of the autonomous network management system of the Field Operational Test Amsterdam. During phase I

of the project the road side measures were successfully designed and operationalized, giving green light to phase 2 where further integration is pursued between in-car systems and the road side infrastructure.

The larger urban regions in the Netherlands started working on the operationalization of network management. To govern the overall transition, a road map of Transport Public Works (2013) was realized for long term future plans ('13-'23) in the policy document 'Beter geïnformeerd op weg' (in English: Better Informed on the road), defining paths and corresponding research questions for the transitions on integrating traffic management measures, in-car developments, traffic data and information, organizations, roles of stakeholders and financing structures.

2.4 Discussion on transition

Designing and operationalizing systems for coordinated and integrated traffic management is a complex task. As can be seen in most of the documents Middelham et al. (1995); Kotsialos et al. (1998); van Kooten & Meurs (2009); Mac Carley et al. (2002) many obstacles are encountered with respect to proper algorithm design, technical implementation, data storage and communication, and conflicting organizational interests. All these programs have in common that it took a lot of effort to get the system up and running.

In this respect, a thorough understanding of control strategies and their impact comes gradually. Literature nicely shows new insights being adopted in the strategies over time, constantly increasing the performance of the control algorithms. Herewith some examples:

- The first coordinated ramp metering schemes that were operationalized within European field operational tests Middelham et al. (1995); Kotsialos et al. (1998) involved multiple simultaneously active local controllers. Upstream ramps were assisting in the metering task, but full utilization of the ramp storage spaces was not targeted. As will be discussed later, this also applied to many of the other initially proposed control strategies;

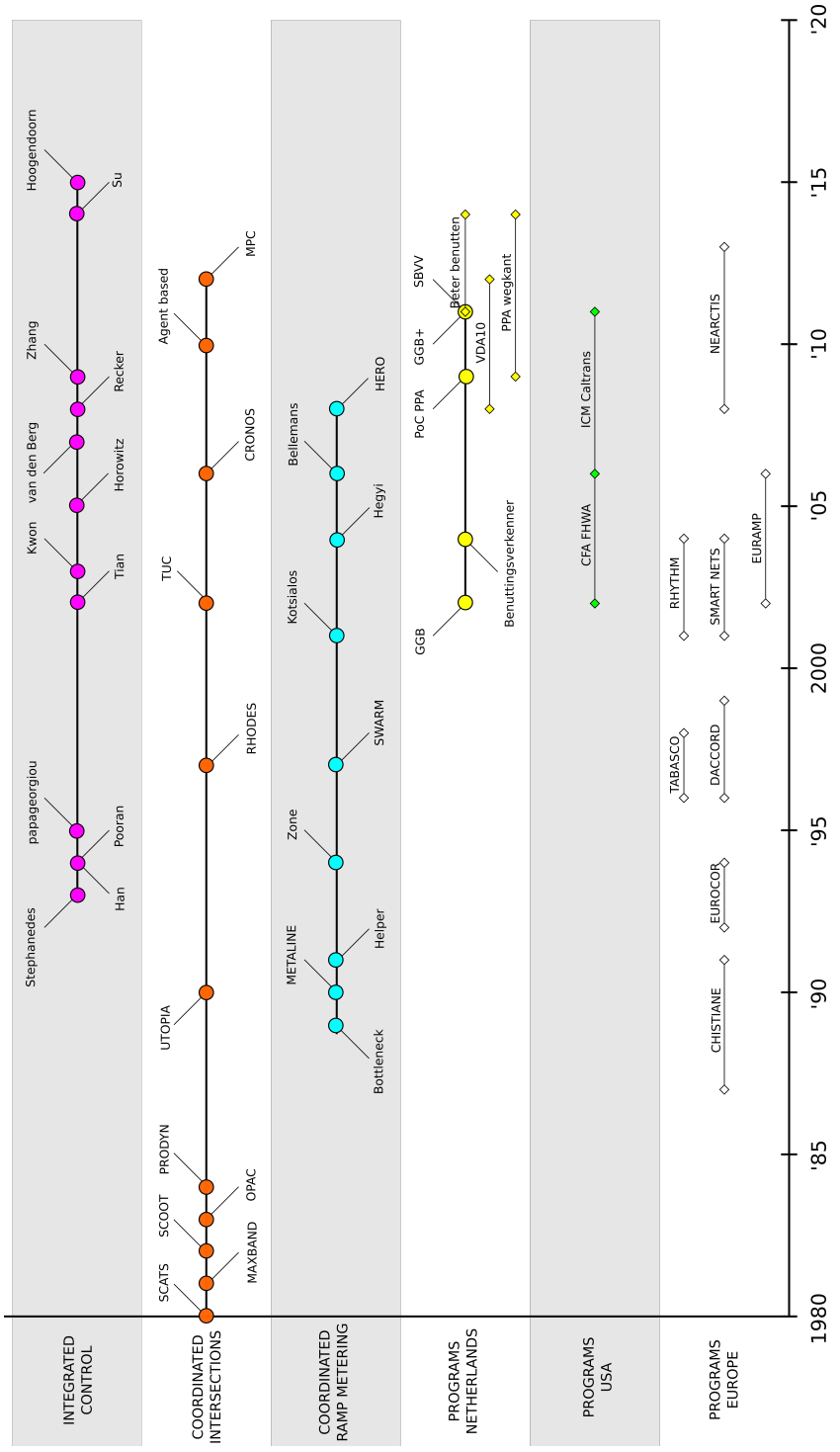

Figure 2.1: Overview of programs, projects and approaches with respect to integrated and coordinated control involving ramp metering and intersection control.

- Subsequently, it became clear that it is important to control the freeway at its true capacity Smaragdis et al. (2004). An algorithm should prevent underutilization as well as over-saturation of the bottleneck at all times. HERO Papamichail & Papageorgiou (2008) was the first coordinated ramp metering scheme to target full utilization of upstream ramp space while explicitly keeping the bottleneck at capacity[1];
- To further increase the efficiently with which upstream ramp space is used, Landman et al. (2015) proposed a coordination strategy that saturates coordinated ramps in downstream order.

There are nevertheless still approaches being operationalized that do not have a clear strategy with respect to improving the network performance. This would not be a problem when aiming at more safety or a cleaner environment, but in the end most evaluations revolve around improving the network performance. In the Dutch field operational test project VDA10 Wang et al. (2009, 2010), the aim was to control the performance of network elements by means of control scenarios in line with the functions and priorities defined in the network vision of the road authorities. To solve traffic problems, solutions were implemented such as reducing a bottleneck's inflow or increasing its outflow. As a result, during oversaturated conditions (i.e. where demand is larger than available supply) the performance of the different roads degraded in line with the priorities without taking the true impact into account of the capacity drop, blocking back and suboptimal use of available road capacity. Moreover, also other aspects of the system make it questionable if the control signals actually lead to better network performance. When operating on a network level, local ramp metering signals could be overwritten by a select number of predefined metering rates; it is unlikely that the freeway capacity becomes fully utilized when applying ramp metering based on three predefined metering modes and where the metering rates are not directly based on prevailing freeway state estimates. The system would also propose solutions that needed to be accepted by the operator to be deployed, which in turn introduced undesirable control delays.

[1] HERO applies a *master-slave concept* in which the master ramp is assigned the task to keep the bottleneck at capacity by means of an adequate local metering algorithm, while upstream slave ramps are assisting in the metering task by reducing their outflow into the mainstream

System design for operational network management should always aim at improved network performance, while satisfying constrains that reflect policy objectives with respect to equity, safety and environment. Therefore, in the design phase of the Field Operational Test Amsterdam there was a strong emphasis on preventing undesired traffic phenomena that decrease the network performance. However, other indicators were introduced to account for policy objectives as: service level agreements on roads and routes in terms of average speeds or flows, maximum storage spaces for locations to constrain the waiting times of stored traffic and agreements on typical situations where the system is directly shut down.

The success of a field operational test strongly depends on the realized network performance benefits, i.e. reducing the total delay caused at the freeway and urban networks that are under control van der Veen et al. (2010); van Kooten & Meurs (2009); Kotsialos et al. (1998). Other objectives are often not given much attention in the evaluation, which implies that the design of the system is normally tailored to meeting the performance objective. This can only be done, if it is truly understood how and to what extend the network performance can be improved, given the other constraints that need to be accounted for.

To conclude, it is important to realize that phenomena such as freeway flow break including its associated capacity drop and congestion spill-back to upstream bifurcation points, can be prevented at the cost of either *holding back traffic elsewhere in the network* or by *rerouting traffic to alternative routes*. Next sections will therefore elaborate on literature with respect to Route guidance, Coordinated intersection control, Coordinated ramp metering and Integrated ramp metering and intersection control.

2.5 Route guidance

In literature many different control approaches can be found for applying dynamic route guidance. This section will give a brief overview of the control approaches along the characteristics feedback versus feedforward and optimal versus non-optimal control to decide upon the desired and required aspects for design.

Feedforward means that the control signals (i.e. the route guidance) are based on the non-controllable inputs (disturbances) of the system, i.e. no correction on the control signals is applied if the real state and desired state are different. Feedback controllers on the contrary do determine control signals based on the system output in terms of up to date measurements or predictions to take unforeseen changes in demand and supply into account. Due to the many uncertainties in practice and the need to systematically control towards a state that reflects the control objectives any proposed control method should desirably be of the feedback type.

Due to their limited complexity, most automated route guidance systems in practice are of the reactive Mammar et al. (1996); Pavlis & Papageorgiou (1999); Minciardi & Gaetani (2001) or predictive Messmer et al. (1998); Wang et al. (2003) feedback control type. They are applied to support road users making optimal routing decisions. Reactive feedback route guidance systems are, however, known to be vulnerable to system oscillations if the impact of a given control signal is delayed. This is the case when the location of the actuator and desired impact are different, and if arrival or instantaneous travel times[2] are used as feedback information Hoogendoorn (1997); Pavlis & Papageorgiou (1999). This system instability can be prevented by using predicted travel times Wang et al. (2001, 2003). If the route guidance method needs to function correctly on a regional level, the impact of control signals based on arrival or instantaneous travel times will be significantly delayed and cause system oscillations. The control signals therefore need to be based on predicted system states.

Optimization-based approaches attempt to optimize a network performance measure by applying a traffic flow model in an iterative optimization procedure. Two types can be distinguished: Optimal Control (OC) and Model Predictive Control (MPC).

[2] Arrival and instantaneous travel times describe the traffic situation respectively before and at the time the route guidance signal is composed. Since traversing a route takes time, traffic situations can significantly change between the moment of giving the rerouting advise and the time the traffic conditions actually change accordingly. If the traffic demand exceeds the available supply, the queues and travel times within both routes can then start oscillating, meaning that queues grow beyond their target length or dissolve completely.

- 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 or even counterproductive. 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. This can be applied in Dynamic Traffic Assignment (DTA) frameworks to analyze system optimal or user equilibrium solutions for large-scale networks Peeta & Mahmassani (1995); Messmer & Papageorgiou (1995) or to realize system optimal signals by optimizing marginal costs Zuurbier et al. (2006); Zuurbier (2010). MPC is also applied to determine optimal route guidance in combination with other DTM measures within urban and freeway networks Karimi et al. (2004); van den Berg et al. (2004).

The MPC control approach is still pending to be implemented in practice due to its complexity and computational requirements. Nevertheless, in an off-line setting, important insights can be gained from the optimal controllers into the key phenomena that cause decreased network performance. This enables the design of much simpler heuristics that achieve comparable results and that can be operationalized.

2.6 Ramp metering

Ramp metering is a powerful means to mitigate the effects of two traffic phenomena that negatively influence the network performance, namely the capacity drop and blocking back of queues. The capacity of a freeway drops with the onset of congestion, because the outflow of a queue is typically 5-15% Hall & Agyemang-Duah (1991); Cassidy & Bertini (1999) lower than the free-flow capacity. Ramp metering algorithms limit the outflow from an on-ramp into the mainstream to prevent a flow breakdown (storing the surplus of traffic with respect to mainstream capacity at the ramp).

Blocking back occurs when a queue spills back to bifurcation points (e.g. off-ramps) upstream in the network where traffic becomes hindered that does not need to pass the bottleneck location Knoop et al. (2007); Papageorgiou & Kotsialos (2002); Landman et al. (2012). By reducing the inflow into the mainstream bottleneck by means of local or coordinated ramp metering, the congestion growth might be slowed down or even stopped.

Two primary categories can be distinguished for operational local ramp metering: fixed-time (or pre-timed) control and adaptive (or traffic-responsive) control. In a fixed-time ramp metering strategy, metering rates are determined based on historical traffic information and established on a time-of-day basis Wattleworth (1965). Due to the absence of real-time measurements they may either lead to overload (congestion) or underutilization of the mainstream capacity. Both situations are undesired when aiming at optimal use of the freeway capacity. The metering rates resulting from traffic-responsive strategies are based on up-to-date state estimates in the vicinity of the ramp. Examples of local traffic-responsive control are feed-forward strategies such as demand-capacity Middelham & Taale (2006) or occupancy control Papageorgiou & Kotsialos (2002), and feedback strategies such as ALINEA and its variants Papageorgiou et al. (1991); Smaragdis et al. (2004).

When applying feed-forward strategies that determine the ramp metering rate based on a freeway state estimate upstream of the merge, it is important that an extra measurement point downstream the merge area is included to adequately respond to the onset of congestion. It must however be noted that empirical research has indicated that the capacity of a freeway system is better expressed in terms of occupancy or density than flow Cassidy & Rudjanakanoknad (2005). Occupancy and density based systems are therefore more adequate in preventing congestion due to under- or overestimation of the actual capacity.

Heuristic ramp metering coordination strategies have been introduced in literature to more effectively prevent congestion at freeways. Some of the first operationalized coordination algorithms include Bottleneck Jacobson et al. (1989), Helper Lipp et al. (1991), Swarm Paesani et al. (1997), METALINE Papageorgiou et al. (1990) and Zone Stephanedes (1994). All of these algorithms include upstream ramps into the coordination in a sequential or parallel manner. However, none of them explicitly targets full

utilization of storage space to maximize the metering duration on a bottleneck. When there remains ramp storage space at the moment the flow breaks down, the metering duration could have been further extended and hence the capacity drop longer prevented. This might explain why evaluation studies do indicate the benefits of applying these coordinated ramp metering strategies compared to local ramp metering Kwon et al. (2001); Hourdakis & Michalopoulos (2002); Chu et al. (2004), but as stated in Zhang & Levinson (2004), it is often not clear how and under which conditions these algorithms improve the overall network performance.

The Helper algorithm has a local ramp metering mode that applies six pre-set metering rates. A ramp that develops an excess queue, increases its metering rate by one step, while the metering rate of the ramp located directly upstream is decreased by one step. If a ramp queue remains critical, the metering rate of the next upstream ramp is decreased a step. This may be repeated until all ramps are assisting. As mentioned before, using upstream ramp storage space to prevent full saturation at more downstream located ramps will extend the metering time on the bottleneck. However, moving a metering task upstream, might not go without the risk of realizing a flow breakdown at the bottleneck location. Notice, that by shifting the metering task to an upstream ramp, the net metering effect on the bottleneck location might become less if not all vehicles from the upstream ramp move along the bottleneck. It should therefore be preferred to only assign the ramp directly upstream the bottleneck the task of preventing a flow breakdown.

The Zone, and Bottleneck algorithms are considering a freeway area with multiple ramps along side. When the area is operating beyond capacity, predetermined weighting factors are applied to distribute the calculated metering task over the coordinated ramps. As long as the applied local ramp metering algorithm prevents a flow breakdown, the coordination scheme makes coordinated ramps store vehicles. However, in the determination of the weights, the on-ramp queue saturation is not explicitly taken into account.

The Swarm algorithm determines excess densities at multiple detector stations and translates these to a volume reduction that is imposed on the involved coordinated ramps by weighting factors based on demand and available storage space. Hence, contrary to Zone and Bottleneck, SWARM ap-

plies weighting factors that include the actual storage space at coordinated ramps. However, it remains unclear if the use of multi-aspect weighting factors actually enables full utilization of storage space at upstream on-ramps. The determined coordination metering rates for upstream ramps should therefore be specifically targeting full utilization of all upstream ramp storage space when metering on a bottleneck.

An approach that does explicitly considers ramp storage space utilization while dealing with a specific bottleneck is HERO Papamichail & Papageorgiou (2008); Papamichail et al. (2010b). The ramp located directly upstream the bottleneck aims at preventing a breakdown. The coordination algorithm activates upstream located ramps to assist in the metering task by synchronizing the saturation degree of the assisting ramps with the saturation degree of the ramp directly upstream the bottleneck. This implies that the ramp that is metering on the bottleneck is likely to run out of space roughly at the same time upstream ramps do, and hence, that not all upstream available space is effectively used to maximize the metering time on the bottleneck. For this reason it might be interesting develop an algorithm that aims at saturating a more upstream located ramp before a more downstream located one, so that all available assistance can be effectively used to extent the metering time on a bottleneck.

By actually targeting maximum use of storage space, these algorithms have the potential to longer or even maximally prevent a freeway flow breakdown with respect to earlier proposed algorithms. However, something that has not yet been explored in literature, is the impact of the set of ramps that are included in the coordination. In other words, can the system performance be further improved by systematically choosing the ramps that need to be adopted into the coordination, such that the combination of resulting ramp and bottleneck delay is minimized.

In literature there are theoretical approaches that determine optimal coordination by means of optimal control Kotsialos et al. (2002, 2001) and model predictive control Papamichail et al. (2010a); Hegyi et al. (2005a); Bellemans et al. (2006). These rather complex methods optimize the metering rates over time by means of a traffic flow model and optimization procedure. However, the complexity and computational demand of those approaches still limit large-scale operationalization in practice. Furthermore, due to non-

linearities it cannot be guaranteed that the global optimal solution is found, and when optimizing over a finite or limited prediction horizon, optimality over the whole simulation horizon is also not guaranteed. Nevertheless, under ideal conditions these methods should be able to determine system optimal ramp metering signals while taking into account aspects such as the optimal set of ramps in the coordination given the prevailing traffic conditions. However, as will be addressed later in this thesis, the optimal ramp configuration is dependent on the peak period duration and the size of the capacity drop. Hence, to adequately determine the optimal set of coordinated ramps, the prediction horizon would need to include the end of the peak period. As mentioned before, long prediction horizons within these complex approaches have a large impact on computational demand. Hence, this identifies the need for a comprehensible approach that can be used to explore the impact of adopting a ramp in the set of coordinated ramps and to gain further insight into this topic.

2.7 Intersection control

2.7.1 Local intersection control

The oldest traffic management measures is probably the intersection controller. They can be classified according characteristics such as fixed time versus traffic-responsive or isolated versus coordinated and integrated. Isolated fixed time controllers were introduced in the 1960s to minimize delays and make better use of the capacity at intersections. The control schemes are optimized off-line based based on typical traffic conditions; this implies that the schemes perform suboptimal in case of unforeseen changes in the demand pattern. Some typical examples are Allsop (1971, 1976); Silcock & Sang (1990); Gallivan & Heydecker (1988). Vehicle responsive strategies make use of real-time measurements that are provided by loop detectors to anticipate on arriving vehicles. In this way, the strategy is able to tailor the control scheme to prevailing conditions and allocate green time more efficiently Vincent & Young (1986); Muller & de Leeuw (2006).

2.7.2 Coordinated intersection control

To further maximize the throughput of multiple controlled intersections within an urban network, Adaptive Traffic Control Systems (ATCS) started to be developed in the late 1970 by optimizing signal splits, cycle times and phase design. Some examples include fixed time based approaches based on off-line optimization methods such as MAXBAND Little et al. (1981) and TRANSYT Robertson (1969), real-time traffic responsive strategies incorporating a network model and optimization procedure such as SCOOT Hunt et al. (1982), SCATS Sims & Dobinson (1980), OPAC Gartner (1983), PRODYN Henry et al. (1984), UTOPIA Vito & Taranto (1990), RHODES Sen & Head (1997) and CRONOS Boillot et al. (2006). These methods generally have large computational demands. Hence, recent developments address more efficient ways for traffic flow modeling and solving the optimization problem, such as the TUC algorithm Diakaki et al. (2002), Agent based approaches de Oliveira & Camponogara (2010) and efficient Model Predictive Control schemes Lin et al. (2012).

2.7.3 Integrated intersection control

These coordinated intersection control approaches are not designed for preventing spill-back of ramp queues to the urban network. There are occasions where, for this reason, the system is taken off-line by its operators Stevanovic (2010). To enable the control of ramp queue spill-back, more research focussed on integrating ramp metering installations with their upstream located intersection controllers.

When designing a system for integrating the control of ramp metering and upstream intersections a thorough understanding is required of the relationship between surface street and freeway operations. In other words, one must understand the phenomena that decrease the network performance such as the capacity drop at the freeway, hindrance to stored vehicles that do not move to the freeway, and spill-back effects on the urban network.

In Han & Reiss (1994) a strategy has been proposed to make more efficient use of the on-ramp space when dealing with non-uniform arrival rates (i.e. platoons of vehicles) each signal cycle of upstream located intersec-

tions. To this aim, the ramp's metering rate is increased when platoons arrive and decreased after they are served, without changing the number of vehicles that can be released to the freeway in a cycle. Despite that the approach minimizes the ramp delay, further increasing the metering rate should essentially result in a breakdown. Note, that if this would not be the case, that the ramp metering installation could potentially release more traffic and is therefore is not keeping the freeway properly at its true capacity.

In Pooran et al. (1994) different strategies have been proposed to handle typical queuing situations at the connection between the freeway and urban network. These strategies aim at preventing *on-ramp queue spill-back to the connection* and *off-ramp queue spill-back to the freeway network*. For instance, if the on-ramp becomes saturated, the ramp's inflow is decreased coming from the feeding intersection arms. When the off-ramp queue spills back to the freeway, the corresponding signal group is prioritized to increase the flow to the urban arterial.

These strategies have been tested by means of simulation Tian et al. (2002), and in Tian et al. (2005); Zhang et al. (2009) dynamic controllers have been proposed to postpone on-ramp and off-ramp saturation by storing vehicles at the arms of the interchange. These dynamic approaches are extremely suitable for implementation. However, one could question the ability of the approaches to adequately stabilize the queue and to use all available storage space. The methods use queue detectors to detect the presence of a queue and subsequently decide on which arms the outflow to the ramp is maximized or minimized. This control approach is rather crude for stabilising a queue and preventing spill-back. Note, that once a direction is prioritized, its queue is dissolved quickly. Such dynamics require a more nuanced control strategy that enables stable and full utilization of intersection buffers.

There are however also approaches that do not even have a clear strategy Kwon et al. (2003); this scheme balances congestion levels between freeway and urban road segments, but gives no argumentation on how and why the strategy improves the network performance. Model based optimization approaches on the other hand should under ideal conditions be able to determine system optimal control signals that minimize the overall delay. Examples of such approaches are Papageorgiou (1995); Stephanedes & Chang (1993); Horowitz et al. (2005); van den Berg et al. (2007); Recker (2003); Su

et al. (2014). As mentioned before, despite the fact that these approaches are able to optimize basically any network configuration, their computational demands are high and their complexity makes employment and interpretation of their strategy to achieve system optimality difficult.

It is surprising that there is only one other operational control approach found in literature that is able to truly synchronize a ramp's inflow with the realized metering rate to keep a ramp queue at constant length Hoogendoorn et al. (2015). This state feedback-based approach stores traffic at intersections upstream located from bottleneck (e.g. oversaturated ramp or intersection) in case the queue threatens to block other flows. All predefined locations to temporarily hold back traffic are simultaneously filled by synchronizing their saturation rate constantly with that of the critical ramp or buffer. This will result in the situation where (under ideal conditions) all buffers run out of space at the same moment. The method is of the reactive type and does not consider (by means of a prediction model and optimization procedure) where and to what extend it is beneficial to store vehicles when dealing with a freeway bottleneck.

For the effective application of this approach, there is a need to a priori define the set of buffers that maximizes the network performance. Moreover, understanding the factors that drive optimal system performance can help in interpreting the behavior of optimal approaches or be a basis for the design of high-performance control strategies that are easier to operationalize and maintain.

2.8 Conclusions from literature survey

In sum, the following research needs are identified as a basis for the work presented in this thesis:

- **In general.** Feedback control is preferred over feedforward control to always ensure the determination of control signals that target a desired network state. Moreover, the duration for the control signal to change the network state (i.e. control delay) needs to be as small as possible to prevent system instability. Correct tuning of the feedback gains is in this respect very important, but still a time consuming task. Optimal approaches have large computational demands and are rather complex

to interpret when applied on large-scale. Hence, heuristics are preferred for operational use as long as they have a clear strategy for preventing bottlenecks from becoming active. Apart from improving the network performance, the parameterisation of the system should also account for other policy objectives.

- **Route guidance.** All of the operationalized automatic route guidance approaches are aiming at user equilibrium conditions, i.e. none of them is aiming at system optimality or taking into account other policy objectives. Route guidance could nevertheless be an interesting means for road authorities to enable network use in line with their policy objectives. To this aim, the objectives can be translated into target service levels (i.e. in terms of travel times, densities, flow, speed within the routes) that are automatically realized by a control algorithm given the prevailing conditions. Due to the large control delays when applying route guidance, predicted state estimates need to be used to keep the control system stable.

- **Coordinated intersection control.** Most coordinated intersection controllers (Advance Traffic Management Systems) optimize the throughput within an urban arterial based on rather complex approaches. The interaction with the freeway system (i.e. preventing spill-back of ramp queues) is not explicitly taken into account. This identifies opportunities for the development of control heuristics that enable using allocated storage spaces in the urban network to longer prevent spill-back from both ramp queues and intersection queues. Moreover, the existing approaches for coordinating ramp metering and intersection control are based on rule based decisions, hence, there are opportunities for designing proper feedback-based control approaches.

- **Coordinated ramp metering.** As previously mentioned, full use of storage space is essential when maximizing the metering duration on a freeway bottleneck. With respect to earlier proposed strategies for coordinated ramp metering, an improvement of the effective use of ramp space is foreseen, if coordinated ramps are saturated in downstream order.

- **Storage space.** None of the heuristic approaches with respect to coordinated ramp metering, coordinated intersection control or integrated ramp metering and intersection control address the issue on the allo-

cation of storage spaces and the utilization of their space. By clearly identifying the relation between system variables (i.e. network, demand and control characteristics) and the costs and benefits of coordination, more efficient control algorithms can be designed that enable further reductions of the total system delay.

- **Control framework.** Most of the proposed control frameworks for integrated traffic control on a network level (i.e. urban and freeway) are optimization-based and not operationalized. With respect to realizing integrated network management in practice, this identifies the need for generic and modular framework design that can be tailored to any congested regional network and implemented stepwise.

Chapter 3

Route guidance in line with policy objectives

Traffic management on a network level is not only theoretically a complex problem, but its practical application also involves the realization of the road authorities' traffic management policy. In the Netherlands this policy harmonizes the interests of involved stakeholders by means of a common vision upon the network functioning, and is expressed in road priorities and corresponding desired service levels. As a first step towards the operationalization of policy into practice, this chapter presents a predictive route guidance methodology that is able to distribute traffic over the network in line with formulated objectives. To this aim, the service levels of routes (reflecting the objectives and priorities) are degraded and restored stepwise. The methodology consists of a finite-state machine that determines the desired service levels based on predicted traffic conditions. These service levels are used by a feedback controller as setpoints for the desired travel times, resulting in the corresponding output signal of a Variable Message Sign.

This chapter is based on work published in:

- Landman, R.L., T. Schreiter, A. Hegyi, J.W.C. van Lint and S.P. Hoogendoorn, Policy-based service level-oriented, route guidance in road networks: a comparison with system and user optimal route guidance, *Transportation Research Record*, Vol. 2278, pp. 115-124, 2012
- Landman, R.L., A. Hegyi, S.P. Hoogendoorn, Service level-oriented route guidance for overlapping routes in road networks: A comparison with MPC, In *Proceedings of the 2012 American Control Conference*, pp. 5775-5782, 2012
- Landman, R.L., A. Hegyi, S.P. Hoogendoorn, Service level-oriented route guidance in road traffic networks, In *Proceedings of the 14th IEEE Conference on Intelligent Transportation systems*, pp. 1120-1125, 2011

3.1 Introduction

Today's increasing adverse effects of congestion indicate the need to apply traffic management on a network level to improve network performance. However, harmonizing the deployment of traffic management measures to operationalize traffic management policies network-wide is complex. Efficient traffic flows are important, but aspects like the environment, livability and safety need to be considered too. To successfully operationalize policy into practice, a control approach is required to systematically steer the network towards the desired state that reflects the policy objectives. Moreover, the control system should produce control actions that are comprehensible for the authorities, because they are the ones responsible for the effects and consequences.

As we have seen in Chapter 2, none of the operational route guidance approaches proposed in literature is suited to deal with the above mentioned desire of realizing network use in line policy objectives. These approaches aim at user equilibrium conditions, meaning that road users are assisted in choosing the route that results in the lowest travel time. By doing so, prevention of phenomena that decrease the network performance or realization of other objectives is not explicitly targeted. Realizing a service level difference between route alternatives could be desirable to prevent traffic problems or to reduce the traffic's impact on the environment or safety. There are approaches available to deal with these objectives, however, they are of the optimization based type. This implies that they are, under ideal conditions, able to maximize system performance or minimize objectives such as emissions. However, operationalization is difficult due to their computational demand and complexity, which identifies the need for a heuristic control approach to distribute traffic.

This chapter presents a heuristic route guidance approach that is able to operationalize formulated policy in a comprehensible and systematic way. The controller is of the feedback type to ensure that the control signals are always targeting the desired traffic conditions within a route, while unforeseen disturbances are accounted for. The approach makes use of (one shot) state predictions to prevent unstable system behavior caused by the delayed impact of a route guidance control signal. A finite-state machine is used

to determine the desired or target service levels within the routes based on predicted traffic conditions. These service levels are used in a feedback controller as setpoints for the desired travel times, resulting in the corresponding output signal of a Variable Message Sign. The method is well scalable, meaning that routes between multiple origins and destinations (even with overlap) can be properly controlled. The proposed approach is thus the first dynamic routing approach that is able to operationalize the Dutch traffic management policy in line with how it is formulated.

By means of a comprehensible test case, the approach is compared with Model Predictive Control (MPC)-based route guidance that realizes system optimal conditions and a feedback controller that realizes user equilibrium conditions. The MPC approach is used to understand the controller's behavior to realize system optimal conditions and to evaluate how well the finite-state machine is able to approach system optimality. Comparing the finite-state machine with the user equilibrium approach, gives insight into the potential network performance improvement with respect to the current state-of-practice on prescriptive route guidance. Results show that the proposed controller is able to prevent or limit the effects of phenomena that cause decreased network production, while also taking the interests of the road user into account.

3.1.1 Common vision upon the network functioning

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 (2002). 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 functions, their priorities and corresponding desired service levels ARANE (2009). In Figure 3.1 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).

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

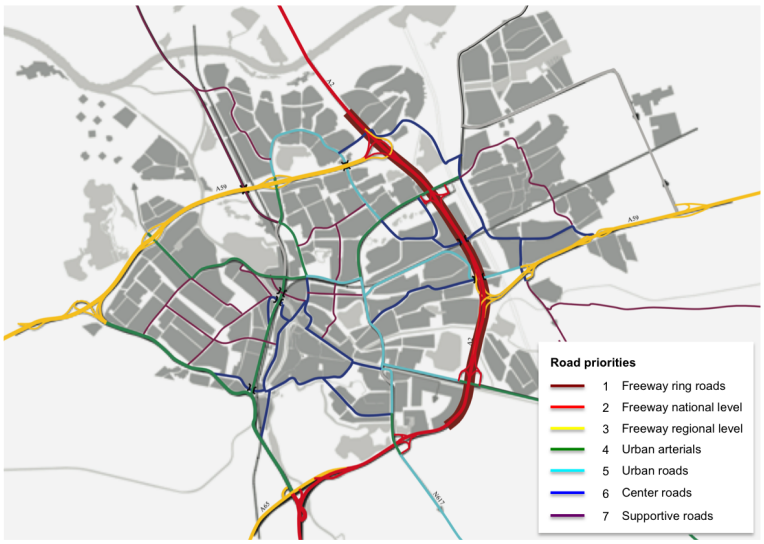

Figure 3.1: Priority map for the city of Den Bosch in the Netherlands indicating with different colors the priorities of the various network elements.

respect considered more important than rural roads, and ring roads more important than arterials. Based on the road priorities and functions, service levels are defined that serve as the foundation upon which to decide where to guide or store traffic. For instance, service levels can be chosen such that:

- Network outflow is maximized by choosing service levels that postpone phenomena as queue spill-back within a route;
- Equity is maintained by setting setting maximum travel time differences between the controlled routes;
- Safety and environmental objectives are realized by imposing flow restrictions within the routes;
- Functional use is operationalized for special scenario's such as events.

3.1.2 Service level definitions

A service level of a network link or route is defined as a performance range, indicated by an upper and lower boundary in terms of traffic speed (or travel time), flow or density. The boundaries determine the acceptable performance

within a service level, and they can differ over the various network elements. The proposed route guidance methodology will be able to use the defined service level boundaries as control targets to establish a desired network state given prevailing conditions. By equally degrading the service levels within the routes, different performance regimes can be established. In the remainder we will identify the time-shortest route (that is generally preferred by most drivers) between an origin and destination as the main route, and all other realistic routes available are considered alternatives.

EXAMPLE: Consider two routes that differ in priority. To protect the outflow of the main route, the alternative is degraded first and with larger steps. Hence, although both routes might have similar free flow speeds, once degraded to their third service level, the main route's minimum performance is 50 km/h and that of the alternative is 30 km/h. A more elaborate discussion on the control of service levels will follow in Section 3.2.2.

In case there are multiple bottlenecks within a route, then indicators such as travel time or average route speed cannot be used to adequately prevent location specific traffic phenomena such as the capacity drop and spill-back. To target such phenomena, it is more evident to temporarily hold back traffic upstream and in the vicinity of the bottleneck. This is also desirable with respect to minimizing control delays, i.e. the time it takes for a control action to have effect. Control delay of route guidance[1] is generally long, meaning that unforeseen disturbances are more likely to occur that nevertheless activate the bottleneck. Coordination strategies to hold back traffic elsewhere in the network will be the subject of upcoming chapters.

The most important phenomena to prevent by means of route guidance are therefore underutilization of available route capacity and limiting congestion growth within routes. The following points are important to account for in the service level definitions with respect to improving the network performance:

[1] The travel time from the route choice location towards the bottleneck determines the time.

- The capacity of the main route should become fully used, before vehicles are sent to time-longer route alternatives. The bottlenecks that determine the route outflows should therefore be activated and released at the same time so that the route set's outflow is maximized;

- As soon as undesired traffic phenomena threaten to occur that decrease the network performance, then traffic can be rerouted to time-longer alternatives. However, the costs of rerouting should not become larger than the saved delays;

- Spill-back of queues (caused by active bottlenecks within a route) should be prevented by allowing the alternative to degrade to some minimum performance that is bounded by a maximum travel time difference over the routes. Hence, the probability of the onset of a traffic problem is prevented within equitable boundaries.

The next section elaborates on the finite-state machine that has been designed to control the service levels of the routes. First the concept is explained for a single route set in Section 3.2. In Sections 3.3 and 3.4 a benchmark is presented to compare the proposed approach with system optimal and user equilibrium route guidance approaches. Sections 3.5, 3.6 and 3.7 elaborate on the scalability of the approach to a network level. Finally, the conclusions and implications are presented in Section 3.8.

3.2 Control approach: Single route set

Route service levels are dynamically controlled by means of a finite-state machine. This is done by a stepwise degradation and recovery scheme that ensures that the performance of the most important route is always better or at least equal to that of the lower prioritized alternative. Based on the prevailing traffic conditions with respect to the predefined service levels, the finite-state machine decides on which route the performance in kept constant, so that the other route is allowed to further degrade or recover.

3.2.1 Control loop for the finite-state machine

In Figure 3.2 the control process is shown. It consists of the following elements: process, model-based prediction, travel time estimation, service levels, finite-state machine and feedback laws. The controller time step is typically larger than the simulation time steps of both the process and prediction model that are used to develop and test our controller. A distinction is therefore made between the simulation time step size T and the time step size T_c after which the controller is activated. This in turn results in the time step counters k and k_c denoting time instants $t = kT$ or $t = k_c T_c$. For the sake of simplicity we assume T is an integer divisor of T_c:

$$T_c = MT, \tag{3.1}$$

with M an integer. When the finite-state machine is activated, the corresponding time index is $k = Mk_c$. The simulation process is modeled by the discrete-time system:

$$x(k+1) = f(x(k), u(k_c), d(k)), \text{ with } Mk_c \leq k < (k_c+1)M, \tag{3.2}$$

with $x(k)$ the state vector of the system (e.g. flow, speed, density) at simulation step k, $u(k_c)$ the control input at control time step k_c (e.g. split fraction), and $d(k)$ the disturbance vector (e.g. demand) at simulation step k.

When the controller is activated the following steps are sequentially executed to determine the control signal for the involved actuators (variable message signs) in practice. State vector $x(k)$ describes the initial network state for a *model-based prediction* that is used to define the future traffic conditions $\hat{\mathbf{x}}(k+1,...,N_p|k)$ over some prediction horizon. Based on this prediction, the *travel times that are to be experienced* $\tau_r(k_c)$ for each route r are determined[2]. In the remainder of the chapter we assume that from a route set $s \in S$ consisting of routes $r \in \{1,2\}$, the main route is always indicated by $r = 1$ and its alternative by $r = 2$. The travel times indicate the current performance of each route, and in combination with the route length

[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 reactive (instantaneous) travel times.

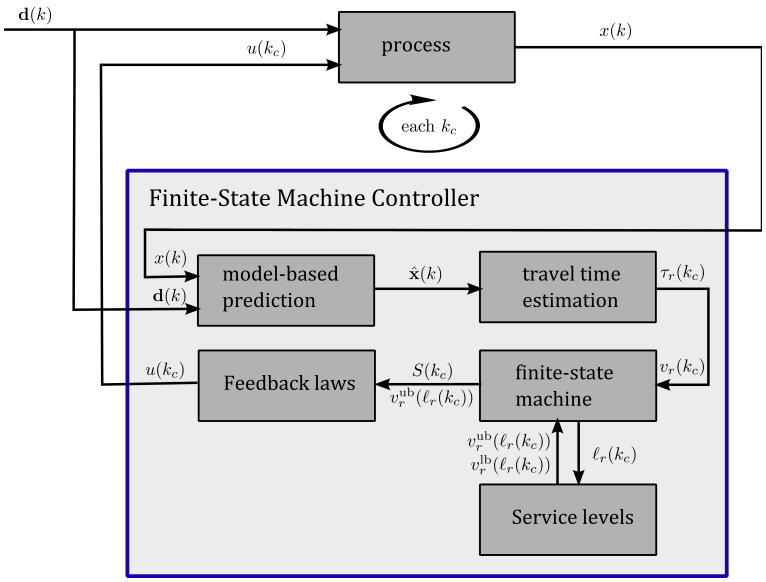

Figure 3.2: The control loop of the finite-state machine.

they are easily translated to the average travel speed $v_r(k_c)$ in km/h. 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.

3.2.2 Control of service levels

The service levels are expressed in terms of travel time or speed, and in Table 3.1 an example is given in terms of speed. For each route, every service level $\ell_r(k_c)$ is determined by an upper boundary $v_r^{\text{ub}}(\ell_r(k_c))$ and lower boundary $v_r^{\text{lb}}(\ell_r(k_c))$. Notice from the table that the boundaries of the same service level can be different for the different routes, and that the level indices increase when the performance degrades.

The functioning of the finite-state machine will be explained by service levels 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

Table 3.1: Service levels $\ell_r(k_c)$ with their upper boundary $v_r^{ub}(\ell_r(k_c))$ and lower boundaries $v_r^{lb}(\ell_r(k_c))$ expressed in km/h.

Level	Main route		Alternative	
ℓ	$v_1^{\text{ub}}(\ell_1(k_c))$	$v_1^{\text{lb}}(\ell_1(k_c))$	$v_2^{\text{ub}}(\ell_2(k_c))$	$v_2^{\text{lb}}(\ell_2(k_c))$
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

prevents unrealistic and unfair travel time differences between route alternatives from being realized and maintained[3].

The finite-state machine updates its state $S(k_c)$ and the active service levels $\ell_r(k_c)$, based on a comparison between the route performance $v_r(k_c)$ and the active service level boundaries $v_r^{\text{ub}}(\ell_r(k_c-1))$ and $v_r^{\text{lb}}(\ell_r(k_c-1))$ from the previous control interval k_c-1. The updated state $S(k_c)$ is used to select and execute the corresponding feedback algorithm. The active service level upper boundary $v_r^{\text{ub}}(\ell_r(k_c))$ serves as setpoint in the state feedback laws to determine the control signal. During this procedure, the finite-state machine encounters either oversaturated or undersaturated traffic conditions:

- **Oversaturated conditions.** The traffic demand to both routes is larger than the joint capacities, resulting in increasing congestion and decreasing performance;
- **Undersaturated conditions.** The traffic demand to both routes is smaller than the joint capacity, resulting in decreasing congestion and increasing performance.

In Figure 3.3 the control process of the finite-state machine is illustrated with the values from Table 3.1. In the example both routes initially perform within their highest service level $\ell_1(0)=1$ and $\ell_2(0)=1$. The performance of the main route is kept constant at the first service level upper boundary $v_1^{\text{ub}}(\ell_1(k_c))$, to make the alternative degrade first during oversaturated conditions.

[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.

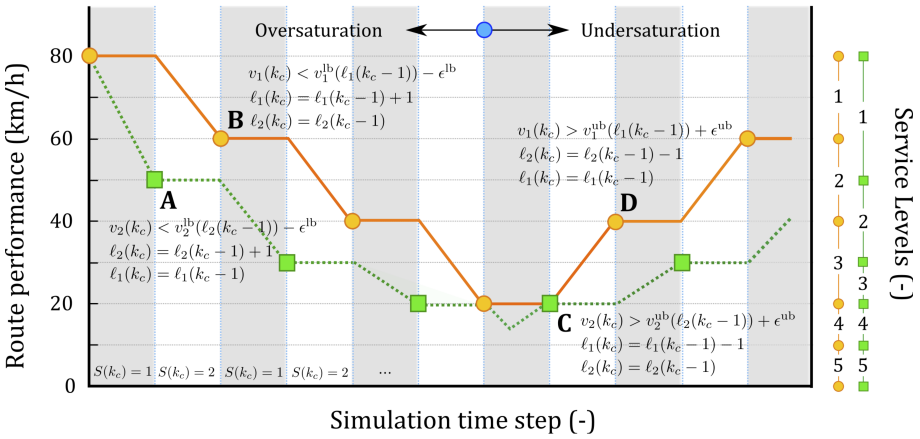

Figure 3.3: Process of service level degradation and recovery, triggered by crossing the specified lower and upper boundaries of active service levels. The orange solid line indicates the main route, and the green dashed line the alternative.

Control process during oversaturated conditions

- The performance at the alternative (green dashed line) will first decrease until its lower boundary $v_2^{\text{lb}}(\ell_2(k_c))$ is reached in point *A*;
- To prevent further degradation in the next control interval, the service level of the alternative is increased to $\ell_2(k_c) = \ell_2(k_c-1)+1$ and the performance kept constant at the upper boundary $v_2^{\text{ub}}(\ell_2(k_c))$ of its second service level;
- The main route is subsequently allowed to degrade until its lower boundary $v_1^{\text{lb}}(\ell_1(k_c))$ of its first service level, which is illustrated by the orange solid line that reaches point *B*;
- In the next control interval, the service level of the main route is increased to $\ell_1(k_c) = \ell_1(k_c-1)+1$ and its corresponding upper boundary maintained;
- If oversaturated conditions remain, this cycle is repeated until all performance levels are used.

Control process during undersaturated conditions

- The performance will increase for the route of which the performance is not kept constant;
- At point C, it can be seen that performance on the alternative does not reach its lower boundary, but moves cross its upper boundary $v_2^{\text{ub}}(\ell_2(k_c))$;
- To enable the main route to recover before the alternative does from the next control interval on, the alternative becomes controlled, and the performance level of the main route is switched back to $\ell_1(k_c) = \ell_1(k_c - 1) - 1$;
- If undersaturated conditions remain, the performance at the main route will improve to point D;
- Then the service level of the alternative is decreased to $\ell_2(k_c) = \ell_2(k_c - 1) - 1$ and the performance at the main route maintained at its active service level upper boundary again.

Preventing oscillations in the switching process

To prevent oscillation in the switching process, an extra threshold is added to the boundaries 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 ε^{lb} and ε^{ub} expressing the threshold as a function of the route length L_r, the considered reference value $v_r^{\text{lb}}(\ell_r(k_c))$ or $v_r^{ub}(\ell_r(k_c))$, and the defined travel time difference μ. The upper and lower boundaries become respectively $v_r^{\text{ub}}(\ell_r(k_c)) + \varepsilon^{\text{ub}}$ and $v_r^{\text{lb}}(\ell_r(k_c)) - \varepsilon^{\text{lb}}$.

3.2.3 The finite-state machine

This switching process can be formulated by means of a finite-state machine that consists of two states $S(k_c) \in \{1,2\}$:

- $S(k_c) = 1$: service level index of the main route and alternative are equal;

- $S(k_c) = 2$: service level index of the alternative is higher[4] than that of main route.

In Figure 3.4 the formal representation of the finite-state machine is given. The finite-state machine states are represented by the squares and on the arrows, triggers can be found that initiate a state transition including the corresponding action of switching the target service levels. The outer loop over the finite-state machine states is followed during the degradation process and the inner loop during the recovery process. If the performance of the route that is allowed to degrade or recover remains within its service level boundaries, no state transition is triggered and the active service levels remain the same. This is indicated by the triggers and actions within each state.

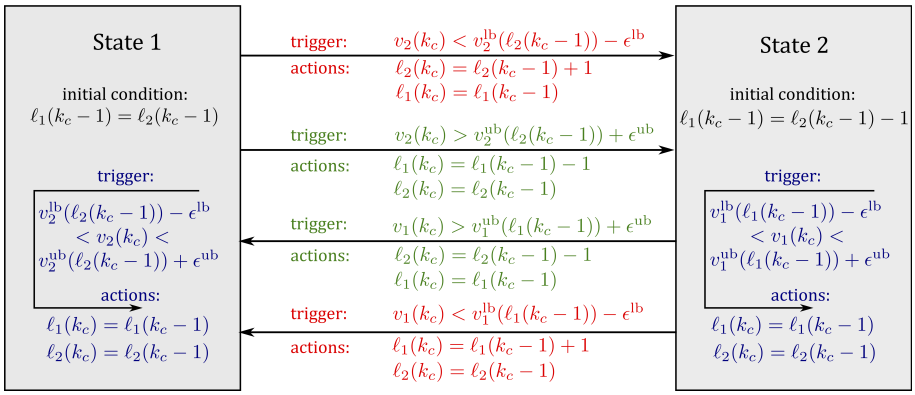

Figure 3.4: Finite-state machine, in the squares the different states of the system, and on the arrows the triggers to make a state transition and the corresponding action of switching the service level.

The feedback control laws given in (3.3) are used to keep the performance constant of the main route and the alternative in respectively states 1 and 2. They determine the desired split fraction $\beta_n^d(k_c)$ (this is the control signal $u(k_c)$ in Figure 3.2) for the controllable traffic flow at the node n directly downstream the VMS 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 frac-

[4]Notice that we refer to the service level indices used in Table 3.1.

tion $\hat{\beta}_n^d(k_c-1)$, a feedback gain α, the current route performance in terms of travel time $\tau_r(k_c)$, and the setpoint $\tau_r^{\text{ub}}(\ell_r(k_c))$ from the service level table in terms of travel time

$$\beta_n^d(k_c) = \begin{cases} \hat{\beta}_n^d(k_c-1) + \alpha(\tau_1(k_c) - \tau_1^{\text{ub}}(\ell_1(k_c))) & \text{if } S(k_c) = 1 \\ \hat{\beta}_n^d(k_c-1) - \alpha(\tau_2(k_c) - \tau_2^{\text{ub}}(\ell_2(k_c))) & \text{if } S(k_c) = 2. \end{cases} \tag{3.3}$$

It has to satisfy $0 \leq \beta_n^d(k_c) \leq 1$, 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 $\beta_n^{\text{N},d}$. The implemented split fraction at time step (k_c) towards the main route then becomes

$$\tilde{\beta}_n^d(k_c) = (1-\gamma)\beta_n^{\text{N},d} + \gamma\hat{\beta}_n^d(k_c), \tag{3.4}$$

and towards the alternative

$$\tilde{\beta}_n^d(k_c) = 1 - \tilde{\beta}_n^d(k_c). \tag{3.5}$$

The compliance γ of traffic towards the control signals defines the response of the controllable flow to a given control signal. It directly determines the resulting network performance since it identifies the boundaries of the achievable control effect (i.e. split fractions) and hence the ability to find a solution under different circumstances. Compliance should therefore be seen separately from the control approach itself.

3.3 Test case: Single route set

The potential of the proposed methodology is illustrated by means of a test case in which it is compared with a user equilibrium feedback controller and the optimal MPC approach. We will show how the finite-state machine switches through the different service levels during over- and undersaturated conditions. The test case further illustrates that by choosing the service levels right, phenomena like blocking back can be delayed or even prevented, and hence the network performance improved. First the applied traffic flow model and the performance indicators are briefly discussed, and then the set-up of the test case, the finite-state machine and the model predictive controller are given.

3.3.1 Applied traffic flow model

The macroscopic multi-class cell-based traffic flow model Fastlane van Lint et al. (2008) has been used for the process simulation, the state predictions of the finite-state machine, and the optimization procedure within the model predictive 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. When queues start blocking important upstream infrastructure, other flows become hindered, causing the network performance to decrease. The flows can be made destination dependent by means of split fraction definitions at the nodes. This enables correct manipulation and propagation of flows that travel between certain origin-destination pairs by means of route guidance.

3.3.2 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 $k = \{0,1,...,K-1\}$ with K the total number of simulation time steps becomes

$$J_{\text{TTS}} = \zeta_1 T \sum_{k=1}^{K} \sum_{m \in M} \sum_{c \in C_m} \rho_{m,c}(k) \lambda_{m,c}, \tag{3.6}$$

with $\rho_{m,c}(k)$ the vehicle densities (in veh/km) over the cells $c \in C$ of all network links $m \in M$, $\lambda_{m,c}$ the corresponding cell lengths (in km) and ζ_1 the functions weight factor. Further, the queue lengths $W_r(k)$ and the realized travel times $\tau_r(k)$ on the routes are evaluated, including an indication of the maximum queue lengths $W_r^{\max}$ per route and the maximum travel time difference $\Delta\tau^{\max}$ over both routes.

The TTS is also used as performance indicator to minimize in the objective function used by the model predictive controller to determine the signals that realize a system optimal distribution of traffic over the routes. The definition is extended with the requirement to limited travel time differences over the complete prediction horizon:

$$J_{\text{MPC}} = J_{\text{TTS}} + \zeta_2 \sum_{s \in S} \max(0, (|\tau_{1,s} - \tau_{2,s}|) - \Delta\tau^{\max}), \tag{3.7}$$

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 $u(k_c)$.

3.3.3 Benchmark with other control approaches

In this section we shortly elaborate on the algorithms with which the proposed approach is compared in the simulation test cases, being a state feedback based user equilibrium route guidance, and model predictive control based system optimal route guidance.

User equilibrium route guidance

It is relatively easy to define a state feedback control law that aims at user equilibrium conditions, meaning that the road users (i.e. with a choice option) are sent towards the route that minimizes their travel time. The controller determines the desired split fraction to the main route $\beta_n^d(k_c)$ for the controllable traffic flow at the node n directly downstream the VMS towards destination d in control interval k_c. The split fraction is a function of the previously applied split fraction $\hat{\beta}_n^d(k_c - 1)$, a feedback gain α and the difference between the (predicted) travel times $\tau_r(k_c)$ of routes $r = \{1, 2\}$:

$$\beta_n^d(k_c) = \hat{\beta}_n^d(k_c - 1) - \alpha(\tau_1(k_c) - \tau_2(k_c)). \tag{3.8}$$

The desired split fraction again needs to be truncated by:

$$\hat{\beta}_n^d(k_c) = \min(\max(0, \beta_n^d(k_c)), 1). \tag{3.9}$$

The actually realized split fractions to the main and alternative route accounting for the compliance rate of traffic are then determined by (3.4) and (3.5). Notice that as the travel time of the main route becomes longer than that of the alternative -resulting in a positive difference in travel times in (3.8)- that the fraction of traffic to the main route is decreased and vice versa.

Model Predictive Control approach

A MPC scheme is used to solve the problem of realizing system optimal route guidance, see Figure 3.5. In MPC, 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 advise given some assumed compliance rate γ. The same procedure is applied as given in (3.4) and (3.5).

3.3.4 Network characteristics

The applied traffic network and its characteristics is shown in Figure 3.6. The VMS to distribute traffic is located in the north. Traffic moves from origin O_1 towards destinations D_1 in the east and D_2 in the south. Destination D_2 can be reached by the main route on the east side or the alternative on the west side. The main route is considered more important since a considerable part consists of a freeway section that is also used by other large traffic flows traveling towards destination D_1. Within each route a bottleneck is located with fixed capacity of 800 veh/h (e.g. representing an intersection or incident) to realize congestion. Traffic is loaded into the network at origin O_1

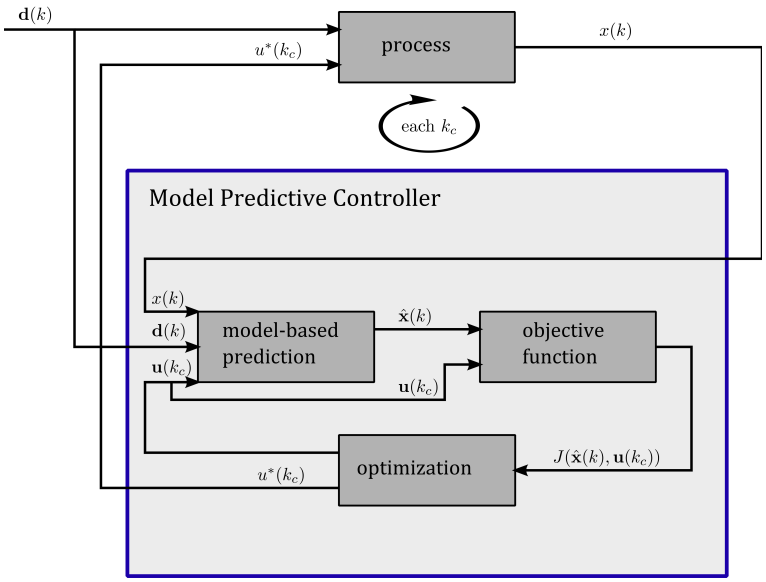

Figure 3.5: The control loop of the MPC approach.

over a three hour simulation period. The inflow at simulation time kT is interpolated from the pattern given in Table 3.5. From the total demand, 50% travels towards destination D_1 and 50% towards destination D_2. The compliance rate γ of traffic to a given advise is assumed to be 30% and the nominal split fraction $\beta_n^{N,d}$ at the node n downstream the VMS towards destination D_2 over the main route is 50%.

Table 3.2: Demand pattern loaded at origin 1.

Time (hh:mm)	8:00	8:30	9:00	9:30	10:00	10:30	11:00	11:30	12:00
Demand (veh/h)	2000	4000	4000	3500	2500	2500	0000	0000	0000

3.3.5 Set-up of the finite-state machine

The policy behind the test case is to increase the network production, with the restriction that the travel time difference over the routes should be less than 10 minutes. The applied service levels are given in Table 3.6. The desired maximum travel time difference is reflected by the maximum travel time difference within a service level (i.e. 10 minutes or 600 seconds). The

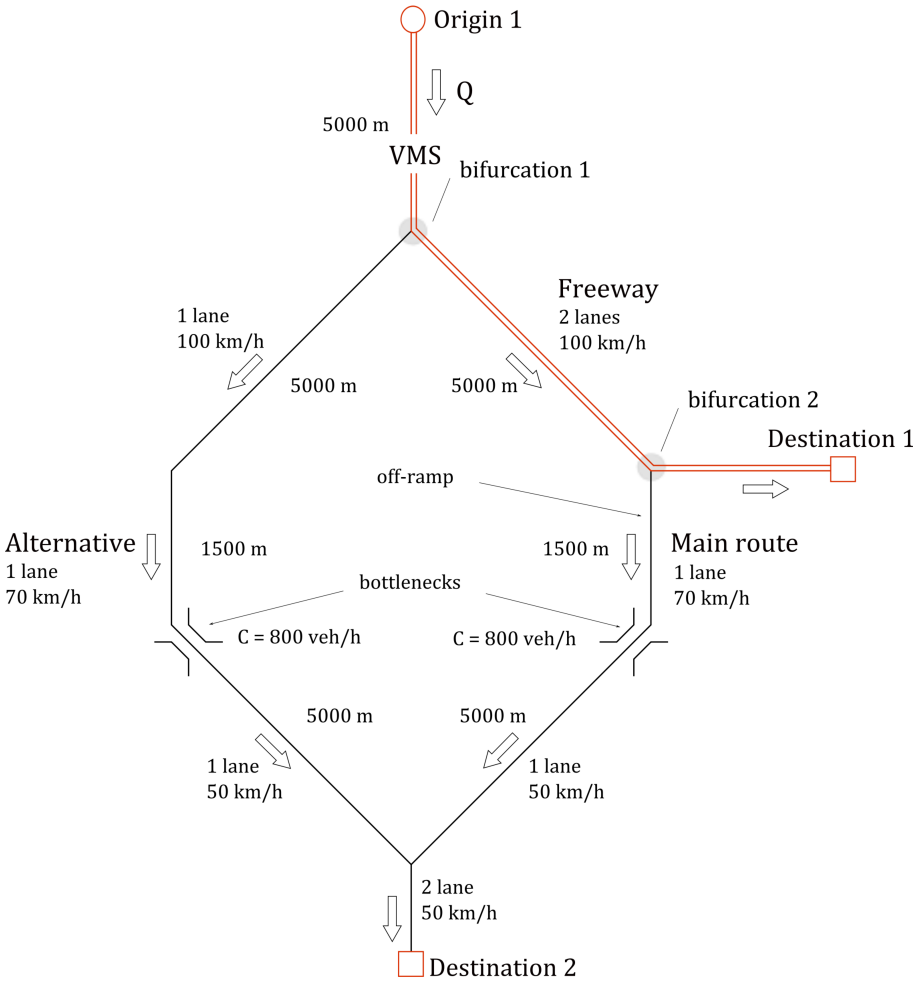

Figure 3.6: Network layout for the test case and its corresponding characteristics. On the east side lies the main route to destination 2, and on the west side its alternative. Ongoing traffic follows the freeway in eastbound direction towards destination 1.

difference per service level is gradually increased towards this maximum value (i.e. 60, 240, 360, 480, 600 seconds for service levels 1 to 7). The service levels per route in this example are degraded by multiples of 60 seconds. The critical travel time at which the congestion in the main route spills back to the freeway is approximately 1200 seconds. Hence, when the alternative degrades to service level 7, blocking back is no longer prevented due to the equity constraint. To conclude, the threshold μ is chosen as 10 seconds and the feedback gain α is chosen as 1.

With respect to the considerations mentioned in Section 3.1.2, by maintaining the first service level upper boundary in the main route, its capacity will become fully utilized before the surplus of traffic is sent to the alternative. Hence, during oversaturated conditions the capacity of both routes becomes fully utilized. Up to the critical travel time of 1200 seconds in the main route, the definition of the service levels is arbitrary; as long as the bottlenecks within the main route and alternative are saturated and no phenomena occur that decrease the network performance, than it does not matter how the queues are distributed over both routes.

Table 3.3: Service level table for the test case. The 1^{st} and 2^{nd} column indicate the service level upper (ub) and lower boundaries (lb) in terms of travel time (s) and the 3^{rd} the corresponding service level in terms of speed (km/h).

levels	Main route			Alternative		
ℓ	$\tau_1^{ub}(\ell_1)$	$\tau_1^{lb}(\ell_1)$	$v_1^{ub}(\ell_1)$	$\tau_2^{ub}(\ell_2)$	$\tau_2^{lb}(\ell_2)$	$v_2^{ub}(\ell_2)$
1	630	690	66	630	690	66
2	690	810	60	690	930	60
3	810	930	51	930	1170	45
4	930	1050	45	1170	1410	35
5	1050	1170	39	1410	1650	29
6	1170	1290	35	1650	1770	25
7	1290	1410	32	1770	1890	23
8	...	...	...	...	...	...

3.3.6 Set-up of the MPC approach

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. If the objective for instance is to minimize the total time spent (related to the outflow of the system), N_p should be typically larger than the maximum travel time from the controlled segments to the exit of the network, because otherwise the effects on the network performance are not accounted for during the optimization. Further, for N_c it is important to find a trade off between the computational effort and the performance of the optimization procedure. To make a choice for an acceptable N_p and N_c both horizons are varied and plotted against the system optimal solutions they return. In that way we can select a combination that guarantees acceptable computation times while still returning the control signals that realize the absolute system optimal solution. The objective function that is used by the controller to realize system optimal control signals is similar to (6.19), with J representing the total time that vehicles have spent in the network.

The controller is tuned by evaluation of the objective function for different combinations of prediction and control horizon. As can be seen in Figure 3.7, there are many different combinations of M, N_p and N_c that approach the absolute system optimal solution for the given network and demand pattern. A prediction horizon of 30 minutes and one variable control signal is already sufficient to realize a system optimum. However, for this test case computational efficiency is not important and we therefore chose parameters that generate a clear control trajectory by extending the prediction horizon and the number of controlled intervals. The implemented prediction and control horizon are set to one hour, divided in 10 variable control intervals of 6 minutes.

3.4 Results: Single route set

In this section the functioning of the different controllers is presented by means of the introduced performance indicators.

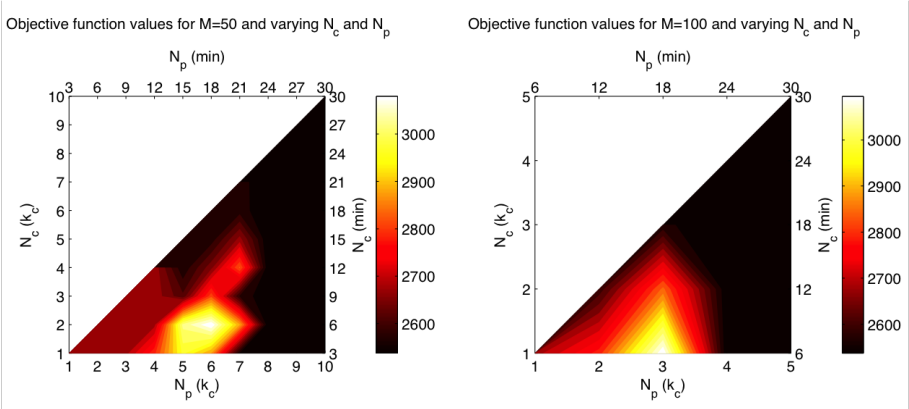

Figure 3.7: System performance for different combinations N_p and N_c.

3.4.1 Control signal MPC

One of the powerful aspects of MPC is that it is able to make the system output and hence the control signals follow a trajectory instead of controlling to a setpoint. To interpret the control signal trajectory it is necessary that it shows a clear pattern. In Figure 3.8D two examples of the optimized MPC control signal over the control intervals are given for the same prediction and control horizon expressed in time, but different control step sizes. Both patterns indicate that from approximately 8:30 to 9:15h a large fraction of the traffic flow is redirected over the alternative, and from 9:45 to 11:00h traffic is guided towards the main route. With respect to achieving system optimality the controller does the following:

- During the oversaturated conditions in the beginning of the simulation both bottlenecks become oversaturated (fully used) so that congestion arises exactly simultaneously at both routes;
- The controller subsequently prevents the queues from spilling back over their upstream bifurcation points. In this way road users who do not have to pass the bottleneck are not hindered;
- Due to the limited storage space on the main route, a large fraction of the traffic flow is rerouted over the alternative, causing a relatively large queue compared to the queue on the main route;

- Since the TTS is dependent on the outflow Hegyi et al. (2005b), the only thing that matters is that both bottlenecks remain oversaturated as long as possible. The allocation of the queues in this respect has no influence on the TTS generated by vehicles traveling towards destination D_2 in the south. If not constrained, this might result in large travel time differences, which is undesired from a policy and operational perspective;

- Another interesting observation is that traffic is directed back to the main route at the moment demand decreases again. Hence, the aim is to keep the bottleneck on the main route saturated, so that its capacity does not become underutilized while there is still congestion on the alternative.

No useful comparison can be made between the control signal of the MPC approach and the finite-state machine. Reason is that the control signal of the latter is determined by realizing the constantly changing setpoints for the feedback controller, while the signal of the MPC approach is determined with respect to realizing system optimal behavior. Nevertheless, the corresponding active service levels of the finite-state machine approach are given in Figure 3.8C to show the stepwise degradation and recovery in terms of service levels.

3.4.2 Travel times and queue lengths

Figure 3.8A shows the travel times as a result of the different control approaches. The stepwise degradation and recovery process from the finite-state machine can be clearly seen, and the travel time difference over the routes remains 10 minutes (see Table 3.4), which is in line with the predefined service levels from Table 3.6. The MPC approach accepts a large travel time difference of more than 17 minutes between the main route and alternative, because it allocates the queues predominantly at the alternative to achieve system optimality. The user equilibrium approach keeps the travel times on both routes equal.

In Figure 3.8B the observed queues for the controlled scenarios are given. The graph of the main route for the finite-state machine shows the delayed spill-back of the off-ramp queue towards the freeway section. This is similar to the strategy of the MPC approach, only less strong due to the requirement

that the travel time difference should not become larger than 10 minutes. The disadvantage of the user equilibrium approach becomes visible. The off-ramp queue from the main route blocks the freeway in an early stage, causing hindrance to the ongoing traffic flow. The queue subsequently grows even faster and this hindrance is the direct cause of the decreased network performance. Figure 3.9 shows an overview of the congestion patterns of the routes that resulted from the different approaches.

3.4.3 Total time spent

The network performance indicators are given in Table 3.4. The TTS is analyzed for the complete system, as well as for the flows towards the different destinations. The TTS of vehicles that travel to destination D_2 are expected to be the same for the different control approaches, because the bottlenecks in the routes determine the outflow. However, if the bottlenecks do not simultaneously become over- and undersaturated, small deviations may be found.

Table 3.4: Overview of the network performance indicators for test case.

	$\mathrm{TTS}_{\mathrm{tot}}$ (h)	$\mathrm{TTS}_{\mathrm{D}_1}$ (h)	$\mathrm{TTS}_{\mathrm{D}_2}$ (h)	$W_1^{\max}$ (m)	$W_2^{\max}$ (m)	$d\tau^{\max}$ (s)
user equilibrium	2784	787	1999	3300	2400	22
finite-state machine	2685	678	2006	2100	3000	616
model predictive control	2660	661	1998	1000	3700	1029

The finite-state machine in this respect performs 0.4% worse than the model predictive controller. This is expected due to the fact that the controller pushes traffic towards the alternative to keep the main route performing at the boundary of its first service level. After the alternative switches to the second service level, congestion will start building up on the main route. Hence, the bottleneck on the main route remains slightly underutilized in the beginning of the simulation. Same reasoning goes for dissolving congestion.

The user equilibrium feedback controller and MPC approach perform the same, since equaling travel times in this test case means that both bottlenecks become over- and undersaturated at the same time. Furthermore, notice that

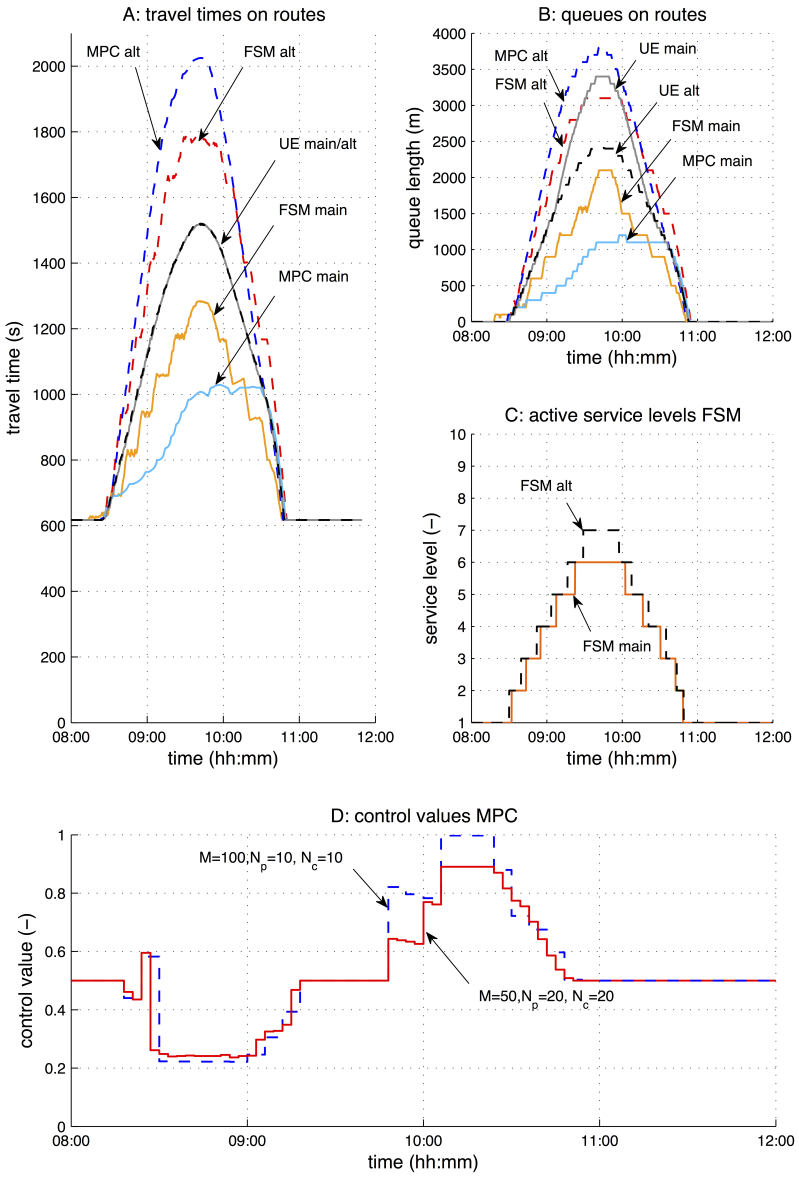

Figure 3.8: Overview of the results with: A. the travel times on the routes resulting from the different controllers, B. the corresponding queues, C. the active service levels of the finite-state machine, and D. the optimal control signals from the MPC approach.

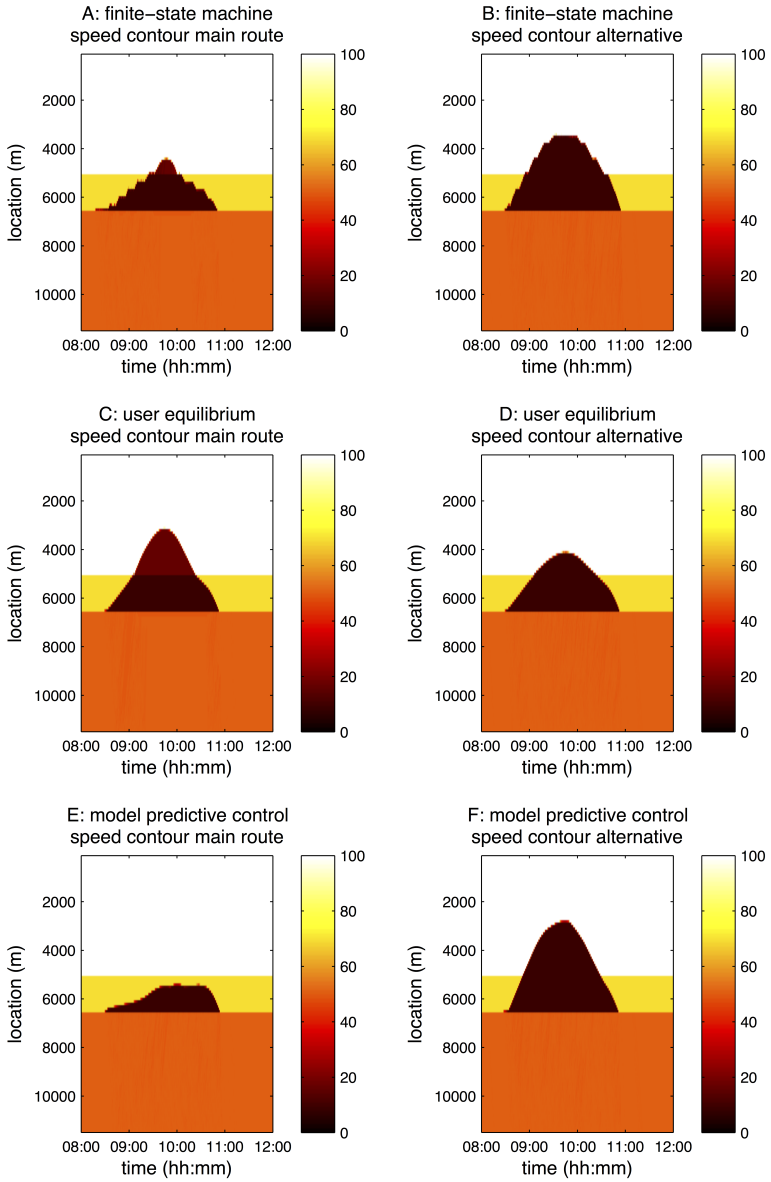

Figure 3.9: Speed contour plots for the main route and alternative as a result of routing by the finite-state machine, the user equilibrium feedback controller, and the MPC approach.

the performance differences between the controllers in this direction are not dependent on the size of the traffic flow towards the south.

With respect to the traffic moving to destination D_1 in the east, the controllers differ significantly in performance. The finite-state machine performs 2.6% worse than the MPC approach. This again, is caused by the chosen policy objective that the travel time difference over both routes should not be larger than 10 minutes. Blocking back of the queue in the main route is therefore delayed, but not completely prevented. The user equilibrium feedback controller even performs 19.1% worse than the system optimal solution in terms of TTS. The finite-state further shows an improvement of 13.9% with respect to the user equilibrium feedback controller (current practice). This performance difference becomes larger if the traffic flow to the east (hindered flow) increases.

3.5 Control approach: Multiple route sets with overlap

The method is also applicable in situations where there are more than two routes between an origin and destination pair, or in situations where routes between different origins and destinations overlap. A simple yet interesting collaboration mechanism between the finite-state machines ensures the utilization of redundant capacity and stepwise performance degradation and recovery among all involved routes. The explanation is supported by the worked example in Figure 3.10.

Routes can overlap with each other by their main or alternative route. It is assumed that the controlled routes by a finite-state machine $r = \{1,2\}$ in route set $s \in S$ initially perform within their first service level $\ell_{r,s}(0) = 1$. Main routes are always indexed $r = 1$ and alternatives $r = 2$. Target service level boundaries then are then defined as $v^{\text{ub}}_{r,s}(\ell_{r,s}(k_c))$ and $v^{\text{lb}}_{r,s}(\ell_{r,s}(k_c))$.

As discussed in previous sections, within a single set of route alternatives, a finite-state machine utilizes available capacity as follows:

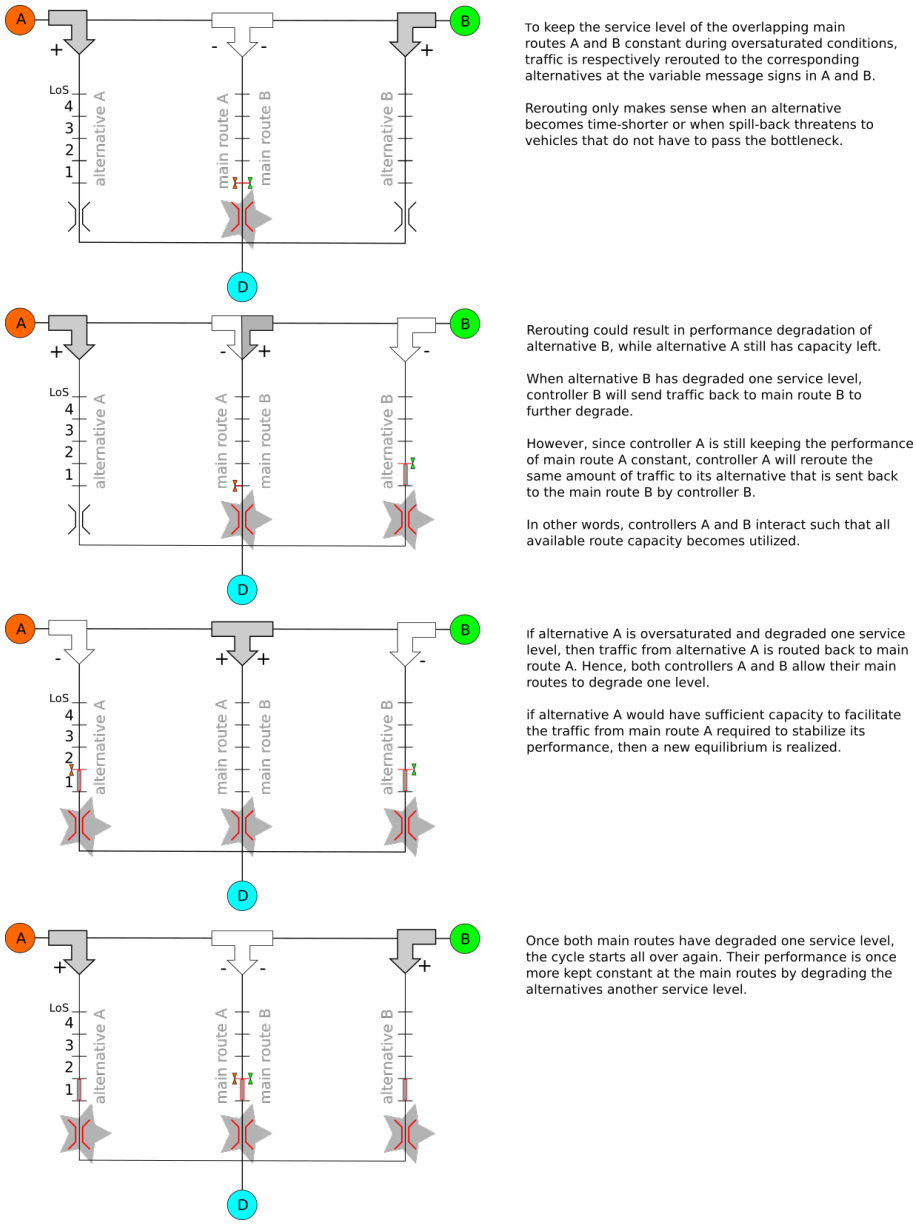

Figure 3.10: Worked example of the interaction mechanism between multiple finite-state machines that distribute traffic over routes that share an overlap, resulting in full utilization of route capacity and degradation in line with target service levels. Grey arrows indicate increasing traffic flow and white arrows indicate decreasing traffic flow. The red brackets identify potential bottlenecks at the downstream end of each route, and the grey stars indicate the bottlenecks that are activated due to oversaturation.

- If the bottleneck impacts a main route, then the controller will directly send traffic to the corresponding alternative, allowing it to degrade one service level;
- If the bottleneck impacts the alternative route, then the controller will start routing traffic to the main route once the alternative degraded one service level;
- If the total amount of traffic can be facilitated by both routes, then the system will find a new equilibrium without further performance decay;
- If the joint capacity of the routes is oversaturated, then the main route and the alternative are degraded stepwise.

Now, imagine the situation where there are two finite-state machines, each controlling the service levels within two routes. Hence, in total there are four routes under control, of which two share an overlap. If rerouting actions of one of the finite-state machines impacts the overlapping part of the routes, then the other finite-state machine will also start rerouting traffic to its available alternative. By allowing one service level difference between routes within a route set, available route capacity becomes fully utilized by the assistance mechanism. The routable amount of traffic, the compliance to a given route advise, and the redundant capacity within the alternatives determine if a problem can be solved (i.e. a route's performance stabilized) and stepwise degradation can be realized.

3.6 Test case: Multiple route sets with overlap

By means of this second test case, the interaction is illustrated between multiple route guidance actuators that are controlled by the proposed approach and compared to a system optimal and user optimal approach. The objective is to improve network outflow, but to keep congestion within the urban network as long as possible. A maximum travel time difference of 3 minutes over the routes is maintained as long as possible. When the queue on the urban network starts to spill back to the freeway, then this equity constraint is relaxed to maximally 14 minutes.

In the remainder of this section, the set-up of the test case, the finite-state machine and the model predictive control approach are discussed. Further, special attention is given to the chosen service level boundaries to realize the above mentioned policy objectives. The applied traffic flow model, the network performance indicators and the used objective function within the Model predictive control approach are similar to that of the previous test case, hence not elaborated on within this section.

3.6.1 Network characteristics

The network for the test case given in Figure 3.11 is inspired by a typical situation within the city of Den Bosch in the Netherlands[5] (see also the priority map in Figure 3.1. Approaching from either the west or east side over the freeway A59 (yellow freeway) there is a main route (middle green radial) and an alternative (outer green radial) available to reach a large event area in the city centre. Hence, there are two route sets of which the main routes overlap.

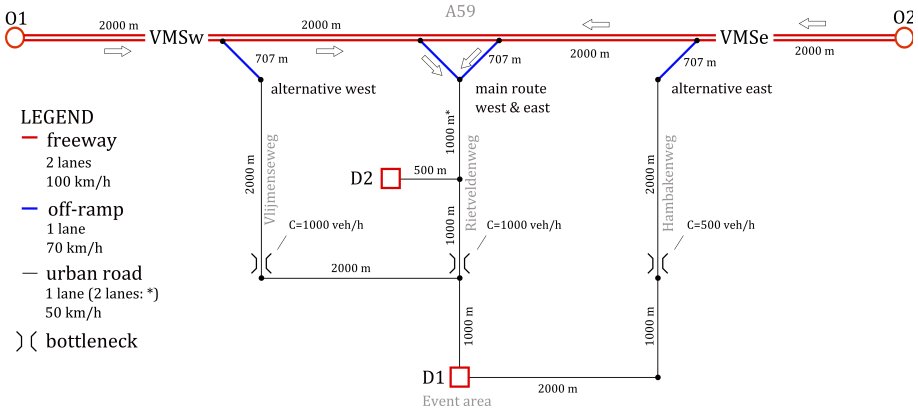

Figure 3.11: Network layout for the test case and its corresponding characteristics.

The test case then becomes as follows. 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

[5]The road authorities of Den Bosch realized a network vision and want to use dynamic traffic management measures to operationalize it.

rates are interpolated from Table 3.5 over a 4 hour simulation period. From both directions, destination D_1 can be reached by a main route in the center 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 250 veh/h. 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 advise 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 3.5: Demand patterns loaded at Origin 1 and 2.

Time	8:00	8:30	9:00	9:30	10:00	10:30	11:00	11:30	12:00
Flow $O_1 - D_1$	0000	1000	1000	1000	1000	1250	1250	0000	0000
Flow $O_2 - D_1$	0000	1000	1000	1500	1500	1500	1500	0000	0000

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. During this phase, the controller will need to use available redundant capacity on both alternatives. From 9:00 to 9:30AM the demand from the O_2 increases and the total demand then equals the total route capacity. Then the controller should find the right distribution of traffic over the routes so that the total capacity becomes fully used. From 10:00AM to 10:30AM the demand from O_1 is increased which causes both route sets to become oversaturated. The quality of the routes then needs to be degraded in such away that the performance difference remains in line with the posed equity constraints (i.e. travel time differences over the routes).

3.6.2 Set-up of the finite-state machine

With respect to the finite-state machine, the applied service levels are given in Table 3.6. 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 the mapping is unambiguous, since each route has only one bottleneck.

- Spill-back within the main routes blocks 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;
- The main and alternative routes' free flow travel times are respectively 396 and 468 seconds;
- The desired degradation step size of 100 seconds results in a maximum travel time difference of 3 minutes between main route and alternative;
- To prevent spill-back to the freeway a maximum travel time difference of almost 14 minutes is assumed acceptable.

These values are used as follows in the definition of the service levels. The free flow travel times over the routes are the upper boundaries of the first service level, and the lower boundaries are acquired by adding the desired degradation step size. The other service levels boundaries follow naturally. When filling in the boundaries, the critical values with respect to spill-back are approached, like the 3^{th} service level upper boundary of the main route (i.e. approaching 600 seconds). To prevent spill-back from blocking the flow to D_2, the 3^{th} service level upper boundary of the main routes is maintained and the alternatives in the west and east are allowed to degrade till congestion reaches their off-ramps at corresponding critical travel time values (i.e. the 3^{th} service level lower boundaries of respectively 1000 and 1400 seconds). Notice that the desired degradation step is relaxed to a maximum of almost 14 minutes. If oversaturated conditions remain, the congestion will be stabilized at the off-ramps of the alternatives, allowing the main routes to degrade till their 3^{th} service level lower boundaries. Hence, it is accepted that the turning direction towards D_2 within the main routes 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. To conclude, the threshold μ is chosen as 10 seconds and the default feedback gain α is chosen 0.001.

3.6.3 Set-up of the MPC approach

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 42 minute prediction horizon of 7 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.

3.7 Results: Multiple route sets with overlap

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 Figures 3.12A, 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 including a short reflection on the applicability in practice.

3.7.1 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. Figures 3.12A-I and B-I show that both controllers directly start redirecting traffic from the main routes to the alternatives to protect the main routes'

Table 3.6: Service level table for the test case. The odd columns of the service level table indicate the upper boundaries $\tau_{r,s}^{ub}(\ell_{r,s}(k_c))$ of the main routes and alternatives and the even columns their lower boundaries $\tau_{r,s}^{lb}(\ell_{r,s}(k_c))$ in terms of travel time (s). Notice that r represents the route index and s the corresponding route set index.

Level	Main route West		Alternative West		Main route East		Alternative East	
ℓ	$\tau_{1,1}^{\text{ub}}(\ell_{1,1})$	$\tau_{1,1}^{\text{lb}}(\ell_{1,1})$	$\tau_{2,1}^{\text{ub}}(\ell_{2,1})$	$\tau_{2,1}^{\text{lb}}(\ell_{2,1})$	$\tau_{1,2}^{\text{ub}}(\ell_{1,2})$	$\tau_{1,2}^{\text{lb}}(\ell_{1,2})$	$\tau_{2,2}^{\text{ub}}(\ell_{2,2})$	$\tau_{2,2}^{\text{lb}}(\ell_{2,2})$
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
6	...	...	...	...	...	...	...	...

performance. Notice that from 8:30AM on only the alternative on the west has redundant capacity left.

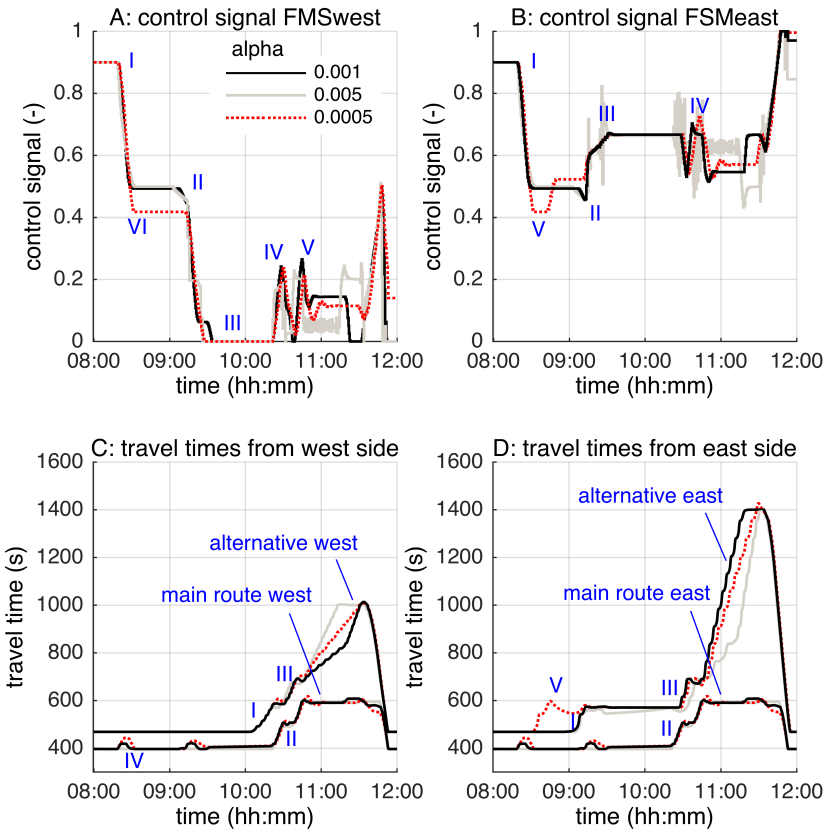

Figure 3.12: Control signals and travel times finite-state machine approach. Notice that a control signal of 1 means that all traffic is sent to the main route.

From 9:00 to 9:30AM the traffic flow from the east increases with 500 vehicles per hour. Figure 3.12D-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 3.12B-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 3.12A-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 Figures 3.12A-III and 3.12C-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 Figures 3.12A-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 3.12D-II. From then on, Figures 3.12A-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 3.12C-III and D-III). Finally, Figures 3.12C 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.

3.7.2 Discussion on tuning the finite-state machines

Without having specified any specific rules for coordinating multiple finite-state machines, traffic is guided such that redundant capacity becomes fully used and that routes degrade stepwise while preventing spill-back in the main routes. The small queue that is initially 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 $\alpha = 0.0005$ by red dashed line in Figure 3.12) 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 Figures 3.12C-IV and D-IV which causes the controllers to send even more traffic to the alternatives indicated at Figures 3.12A-VI and B-V. This is no problem for alternatives with sufficient redundant capacity as can be seen in Figure 3.12C 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 3.12D-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 cause the signal to oscillate as can be seen in Figures 3.12A and B by the grey continuous line of $\alpha = 0.005$. However, even though signal oscillations are undesired from an application point of view, they do realize the desired system behavior because the controllers adequately deal 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.

3.7.3 Model predictive control 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 3.13A-I and B-I, before sending traffic to the alternatives. At first, all traffic from the west is sent to the alternative, however, Figure 3.13A-II shows that part is sent back to keep on utilizing the main routes' full capacity. Main routes are the time-shortest and 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 D2 not hindered. Figure 3.13C 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 3.13A-IV), compensated by sending (slightly more) traffic from the eastern main route to its alternative (see Figure 3.13B-II). Congestion quickly grows on the alternatives until it reaches the off ramps of the freeway (see Figure 3.13C-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 3.13C-II and D-II). The reason for the controller to accept the direction towards D_2 to become shortly blocked at Figure 3.13C-III and D-III might be due to the requirement to release all bottlenecks at the same time.

The optimal strategy can be distilled from the control trajectory, serving to gain insight into properly choosing the service level values such that the finite-state machine mimics the optimal strategy. The MPC approach in that sense can be used to determine adequate control actions for typical situations that regularly occur within the network.

3.7.4 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 3.14A and B. In Figure 3.14A-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 fast increase of travel time on low capacity alternatives, leads to fast increase of congestion on the main routes. As these city arterials distribute traffic over the urban area, flows that do not need to pass the bottleneck become easily hindered (like turning flow towards D_2). This in turn has a negative impact on the network performance. In Figure 3.14C and

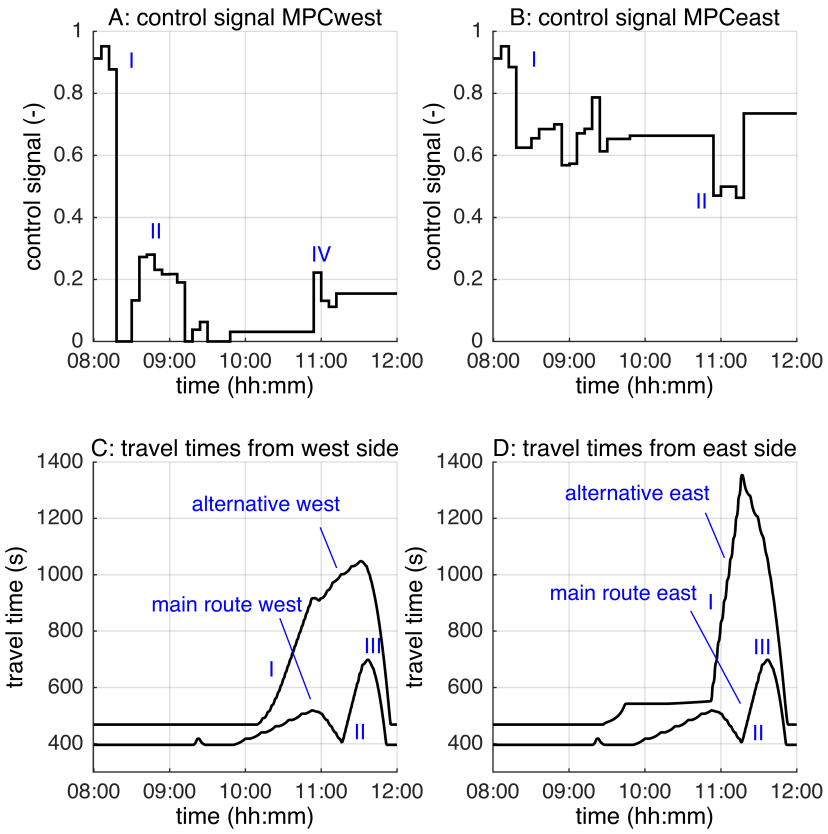

Figure 3.13: Control signals and travel times MPC approach.

D at 10:30AM it can be seen that the travel times on the main routes exceed their critical value (600 seconds, see Section 3.6.2) and cause hindrance to the turning flow towards D_2. The other approaches prevent this hindrance.

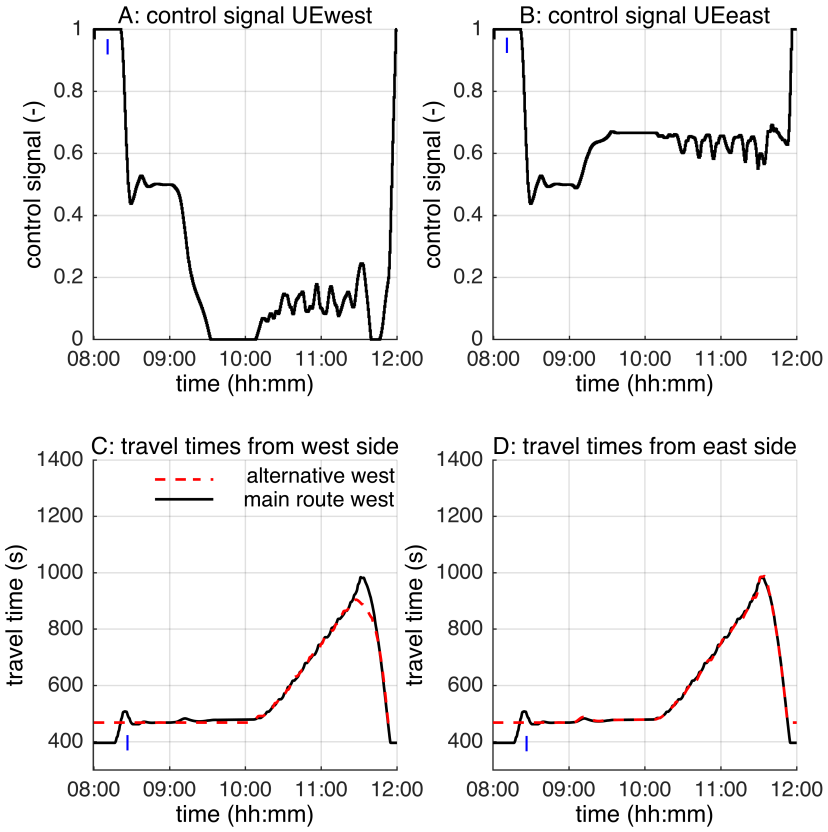

Figure 3.14: Control signals and travel times user equilibrium approach.

3.7.5 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 3.15A. 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 cpu-times 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 cpu-time 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 3.15B shows this problem when we increase the length of the control horizon from 1 to 7 per actuator.

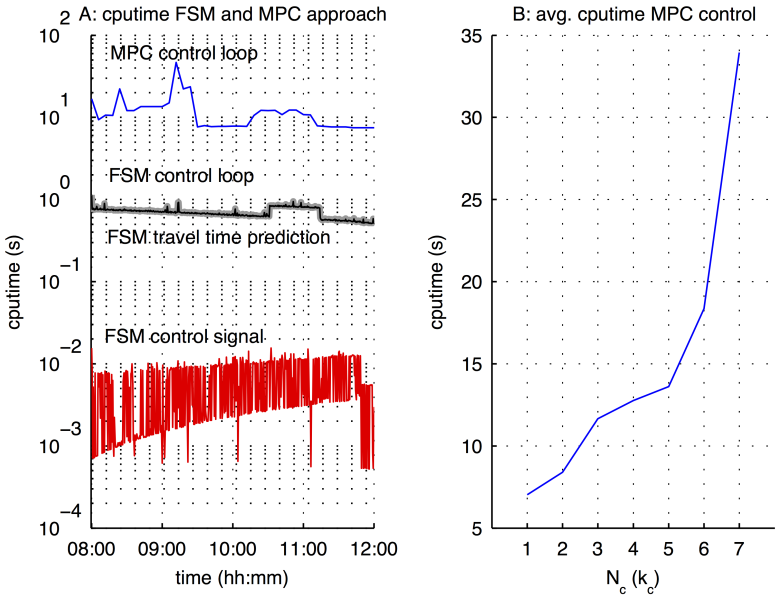

Figure 3.15: Computational demand of finite-state machine and MPC approach.

3.7.6 Network performance indicators

In this section the total time spent by vehicles in the network per destination and the maximum queue lengths per route are given in Table 3.7 for all three approaches. 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 a little less efficient, because it cannot adequately anticipate on this matter. The interaction mechanism requires that one of the alternatives degrades one service level before all available capacity becomes utilized. The user equilibrium feedback approach performs worst - before redundant capacity is utilized, the main routes are degraded till their travel times are equal to that of the time-longer alternatives. This leads to underutilization of available capacity during the degradation and recovery process.

Equalizing the travel times over all routes also causes the most hindrance to turning traffic to D_2. The finite-state machine performs most favorable for traffic moving to D_2, since it prevents the turning direction to become blocked. The optimal approach does allow the turning direction to become shortly blocked (see also Section 3.7.3), but nevertheless has the best performance. This is also indicated by the maximum queues detected on the main routes $W^{\max}_{\text{mWest}}$ and $W^{\max}_{\text{mEast}}$.

Table 3.7: Overview of the network performance indicators TTS and maximum queue lengths for the test case.

	TTS_{D_1} (h)	TTS_{D_2} (h)	TTS_{TOT} (h)	$W^{\max}_{\text{aWest}}$ (m)	$W^{\max}_{\text{mWest}}$ (m)	$W^{\max}_{\text{mEast}}$ (m)	$W^{\max}_{\text{aEast}}$ (m)
finite-state machine	1404	215	1619	2483	1000	1000	1372
user equilibrium	1440	259	1699	1900	2000	2000	739
model predictive control	1386	218	1604	2600	1300	1300	1267

3.8 Conclusions

In this chapter we have presented a service level based route guidance approach that is able to route traffic in line with the traffic management policy objectives of the road authorities in a comprehensible and systematic way. The routes under control are degraded and recovered stepwise by using predefined service levels to realize desired system states that reflect the objectives.

The method is able to improve the network performance during undersaturated and oversaturated traffic conditions by targeting full use of redundant route capacity and prevention of the spill-back phenomenon. This also holds for the situation where multiple finite-state machines are applied and where the controlled routes share an overlap. During oversaturated conditions, the onset probability of phenomena such as spill-back within a route can be reduced, or even prevented in case there is a single bottleneck responsible for the delay. As illustrated by means of the test cases, the finite-state machine is able to realize significant improvements in network performance compared to the user equilibrium feedback controller. System optimality can be even approached if service level definitions allow replicating the system optimal strategy.

Service level based routing (in terms of travel time) does not guarantee unambiguous handling of the blocking back phenomenon when routes have multiple bottlenecks that cause delay[6]. This is likely to be the case in more complex and large-scale networks with multiple route sets between origins and destinations. Degrading the performance of high priority routes in smaller steps than their alternative will at least postpone the moment that phenomena such as blocking back occur. To adequately prevent problems under more complex conditions, the application of coordinated and integrated control approaches is preferred such as local and coordinated ramp metering and coordinated intersection control. Hence, these control approaches are able to temporarily hold back vehicles in the vicinity of the

[6]The test cases illustrate the control approach for comprehensible cases in which route travel times are influenced by a single bottleneck.

bottleneck to prevent location specific traffic problems such as queue spill-back or the capacity drop at a freeway stretch.

With respect to network-wide implementation of the considered routing approaches, we can make the following remarks concerning computational demand, scalability and comprehensibility. Computational time of model predictive control based approaches typically increases exponentially with respect to the number of control signals that need to be optimized. Our test-case shows the same relation. This severely limits the applicability of MPC on large-scale in practice. The finite-state machine approach on the contrary does not have such high computational demands. Its decision making logic is relatively simple and it only needs a single modelrun to determine the predicted travel times for the vehicles that are to be routed. Other aspects that make the FSM approach applicable in practice are the facts that: the method is designed in light of the ongoing developments in the Netherlands (service level definitions for regional networks) and the realized control signals are comprehensible and easy to interpret.

An important prerequisite for good results is that the policy objectives (often qualitatively defined), the priorities of the routes and their functional requirements are carefully translated into maintainable service level boundaries. The boundaries can be chosen such that the controller realizes conditions either closer to system optimal conditions or user equilibrium conditions. The ability to set a maximum travel time difference when defining the service level boundaries ensures that the travel time differences will remain acceptable between route alternatives. The finite-state machine can also be tuned such that it realizes environmental or safety related policy objectives, by means of target service levels in terms of flow or density or accumulation.

Chapter 4

Storage space allocation and utilization

When applying ramp metering to keep a freeway bottleneck at capacity, on-ramp saturation can be postponed by storing vehicles that move towards the bottleneck at upstream located on-ramps along the freeway or at upstream located intersections within the urban arterial. Whether it is beneficial to include a buffer into the coordination, depends on the network characteristics that determine the costs of storing traffic and the benefits of postponing the considered phenomenon that will decrease the network performance. In this chapter an evaluation approach is put forward to make a decision on which buffers to coordinate. To quantify the involved delays, cumulative inflow and outflow curves are developed as a function of the involved system variables such as demand profiles, turn fractions and network layout. This enables us, on the one hand, to determine a priori the optimal set of storage spaces to coordinate and, on the other hand, to gain insight into how the system variables impact the system delay.

This chapter is based on work published in:

- Landman, R.L., A. Hegyi, S.P. Hoogendoorn, Urban Storage Space Selection Method for Integrated Control on a Freeway Bottleneck, *Transportation Research Record*, Vol. 2554, pp. 77-88, 2016
- Landman, R.L., A. Hegyi, S.P. Hoogendoorn, On-ramp Selection Methodology for Coordinated Ramp Metering Schemes, In *Proceedings of the 2015 IEEE Conference on Intelligent Transportation Systems and Control*, pp. 1129-1136, 2015.

4.1 Introduction

The onset of traffic phenomena that decrease the network performance, such as the capacity drop and spill-back, can be postponed (or even prevented) by reducing the flow of vehicles towards the corresponding bottleneck. In the previous Chapter 3, it has been shown and discussed that this can be done by means of route guidance. However, the application of route guidance is better suited to realize an average quality of service within a route, than to prevent the activation of location specific bottlenecks within it.

A more adequate way to deal with location specific traffic phenomena, is to reduce the bottleneck's inflow by temporarily storing vehicles - that are moving towards the bottleneck - elsewhere in the network, preferably near the bottleneck. Holding back vehicles at, for instance, upstream-located on-ramps or intersections can be done by realizing coordination between available traffic controllers.

The duration with which a problem can be prevented by means of coordinated measures is directly related to the amount of available storage space and the efficiency with which the buffers can be used. It is in this respect important to realize, that when vehicles traveling to the bottleneck are temporarily stored elsewhere, also vehicles are hindered that do not need to travel past the bottleneck. Hence, the lower the buffer's fraction of traffic towards the bottleneck, *the more* vehicles need to be held back to realize a certain outflow reduction and *the less* efficiently the storage space is used to postpone a freeway flow breakdown or blocking back of a queue. In other words, including a storage place in the coordination with a low fraction of vehicles towards the bottleneck, leads to a relatively high amount of storage delay, while the effect on metering duration extension and the bottleneck delay reduction remain limited.

In literature there has been significant attention for the development of coordinated control approaches that generate network performance benefits by reducing bottleneck delay at the cost of storing vehicles elsewhere in the network. However, no research specifically addresses which buffers to in-

clude into the coordination and how their space is best utilized with respect to the resulting system delay.

It therefore remains unknown and not thoroughly understood, whether the benefits of preventing a freeway breakdown remain larger than the costs of storing vehicles elsewhere in the network when applying coordinated control. Moreover, most coordination approaches do not consider a clear strategy with respect to the use of available storage space in relation to minimization of delays. Hence, it also remains unclear to what extent an approach improves the network performance, and if further delay reductions can be achieved by choosing different buffers or a different space utilization strategy.

In this chapter an assessment approach is proposed to determine which storage spaces to include in the coordination that minimize the total system delay given the conditions at hand. To quantify the total system delay, cumulative inflow and outflow curves are defined for all locations in the network where delays are caused as a function of involved system variables. In other words, the impact of the key-coordination mechanism on the system delay is modeled by means of these curves. As can be seen in Figure 4.2 the considered locations are: intersection buffers (or arms) upstream of a ramp, on-ramps, and the freeway bottleneck.

The storage space allocation approach is applicable to two forms of coordinated control that are able to postpone a ramp's saturation when ramp metering on a freeway bottleneck. Increasing the metering duration can essentially be achieved by:

- **Increasing the ramp's outflow.** Coordinated ramp metering schemes enable the use of upstream-located ramp storage space, such that the ramp that is keeping the bottleneck at capacity can increase its outflow without causing congestion. To this aim, upstream ramps included in the coordination are assisting in the metering task by reducing their flow to the freeway. The ramp directly upstream the bottleneck can subsequently benefit from the reduced mainstream flow, because it will fill at a slower rate when releasing more traffic from its ramp into the bottleneck without causing congestion;

- **Decreasing the ramp's inflow.** Another way of postponing the moment the ramp becomes saturated is by realizing coordination between the ramp metering installation and the directly upstream-located intersection controller that feeds the ramp. Spill-back of the ramp queue is then prevented by holding back the vehicles that travel to the ramp at its upstream-located intersection arms.

A schematic overview of both applications is given in Figure 4.1. In the coordinated ramp metering case, the considered storage spaces consist of the ramps that are included into the coordinated ramp metering scheme. In the other case, the considered storage spaces are the intersection buffers (or arms) that are located directly upstream a ramp. Both applications can be combined into a single approach, however, this is outside the scope of this chapter.

The structure of this chapter is as follows. In Section 4.2 we will first introduce the typical situation in which the above mentioned forms of coordination are applied in practice. Sections 4.3 and 4.4 and 4.5 elaborate on how the cumulative curves are prepared and the delays quantified for key locations in the network. The storage space selection approach is illustrated in Section 4.6 for coordinated ramp metering and in Section 4.7 for integrated ramp metering and intersection control. In Sections 4.8 and 4.9 the approach is used to evaluate storage space filling strategies such as parallel filling and sequential filling of buffers. The conclusions are finally summarized in Section 4.10.

4.2 Network characteristics and assumptions

In this section the situation and the system variables are introduced with respect to applying the approach for either coordinated ramp metering or integrated ramp metering and intersection control. As can be seen in Figure 4.2, the network can consist of multiple on-ramps $r \in R$ (numbered in upstream direction) including their directly upstream located intersection arms $b \in B_r$ that feed ramp r with traffic that is moving towards the freeway.

- To reduce the flow into a freeway bottleneck, vehicles can be stored at on-ramps $r \in R^c$ located upstream of the bottleneck by means of local or coordinated ramp metering. The controlled ramps R^c are a subset of

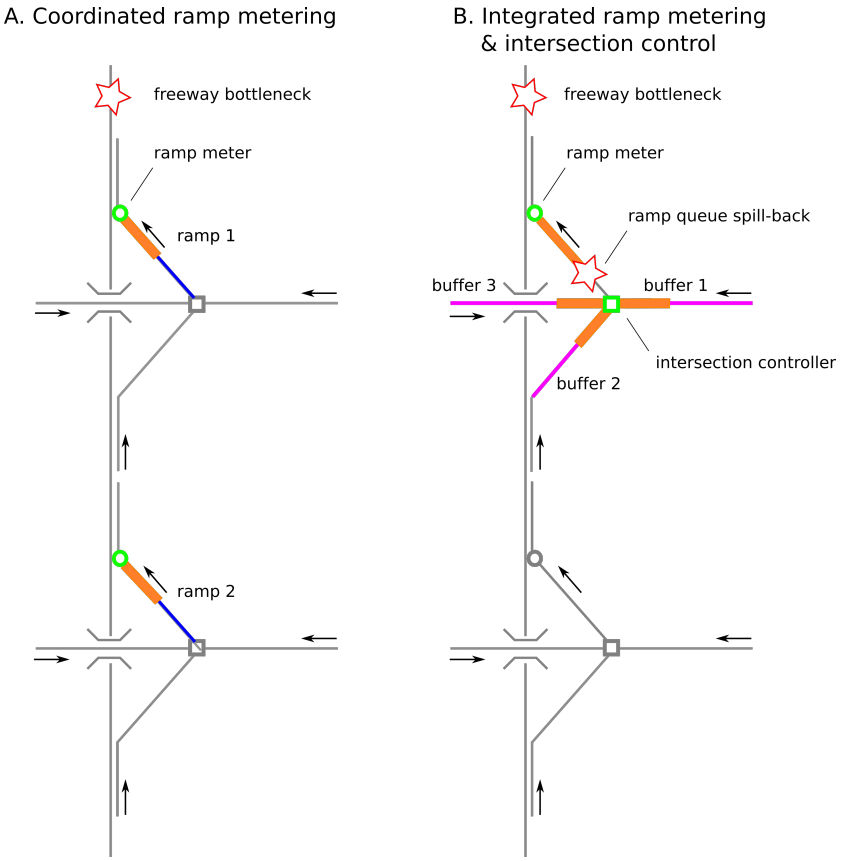

Figure 4.1: The configurations for which the approach can be applied with (a) coordinated ramp metering and (b) integrated ramp metering and intersection control.

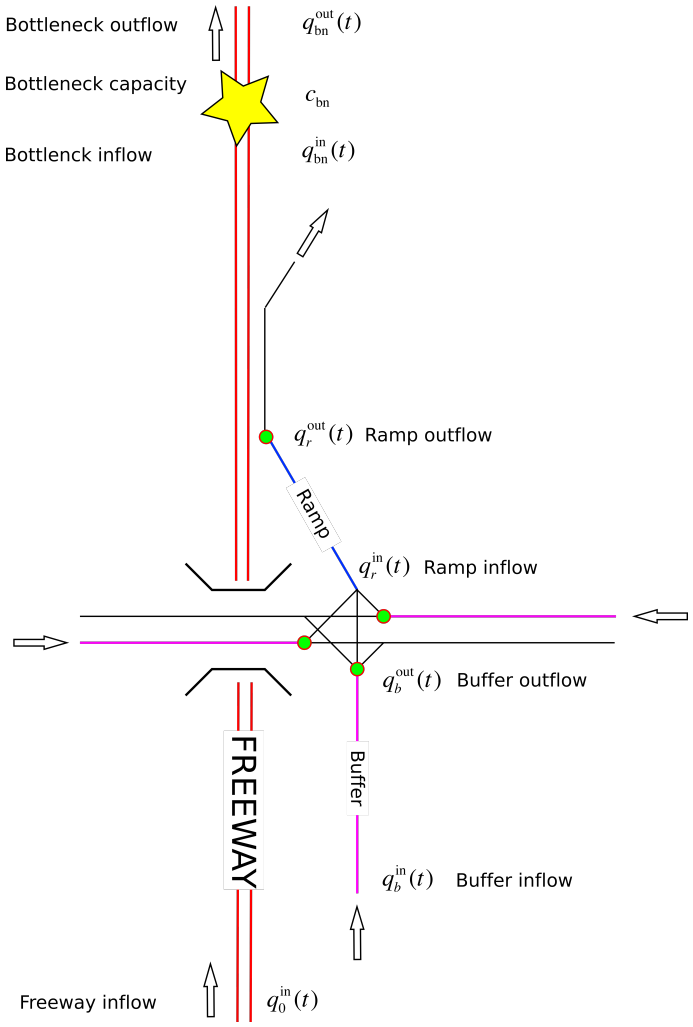

Figure 4.2: The network situation for (a) the integrated control of a ramp meter and its directly upstream-located intersection controller, and (b) the coordinated control of ramp meters in case there are multiple such connections located along side the considered freeway stretch.

all ramps R located along the considered freeway stretch, i.e. $R^{\mathrm{c}} \subseteq R$. The storage capacity of the ramps r is indicated by s_r in terms of vehicles;

- To reduce the flow into an on-ramp r, vehicles can also be temporarily stored in the coordinated buffers $b \in B_r^{\mathrm{c}}$ at the intersection arms located upstream of the ramp. The buffers B_r^{c} used during this form of coordination between the ramp and intersection controller are a subset of all the intersection arms B_r that feed ramp r, i.e. $B_r^{\mathrm{c}} \subseteq B_r$. The storage capacity of the intersection buffers b is indicated by s_b in terms of vehicles.

Traffic flows in the network at respectively the freeway and intersection origins with rates $q_0^{\mathrm{in}}(t)$ and $q_b^{\mathrm{in}}(t)$ in terms of vehicles per hour. Destination dependency of the flows is taken into account by means of traffic fractions, indicating the percentage of traffic moving from a certain origin to a certain destination. The fractions will be elaborately discussed in the next sections. A peak period and an off-peak period are defined in which the demands for every origin are constant over time. The ramp and intersection buffer inflows and outflows over time are represented by respectively $q_r^{\mathrm{in}}(t)$ and $q_r^{\mathrm{out}}(t)$, and $q_b^{\mathrm{in}}(t)$ and $q_b^{\mathrm{out}}(t)$. Downstream the freeway a bottleneck is located with capacity c_{bn}. Its inflow and outflow over time are described by respectively $q_{\mathrm{bn}}^{\mathrm{in}}(t)$ and $q_{\mathrm{bn}}^{\mathrm{out}}(t)$. Moreover, the following assumptions have been made:

- At $t = 0$ the bottleneck becomes oversaturated (start peak period) and the ramp outflows then need to be reduced to prevent the capacity drop;
- Inflows and outflows remain constant during typical process phases, i.e. only at a phase transition the flows change. Examples of such phases are: the peak period in which the bottleneck is over-saturated, the phase where the coordination is active to longer prevent a flow breakdown, the phase where ramp metering rates are increased to prevent further spill-back, and the remaining off-peak period where freeway congestion is able to dissolve;
- When storage space runs out during the peak period, both the ramp and buffers will maximize their outflow to prevent further spill-back of queues by realizing a predefined *flush metering rate*, resulting in the on-set of freeway congestion at the bottleneck;

- If queues can dissolve due to undersaturated conditions at the freeway, then the maximum ramp outflow is constrained by the ramp capacity or the acceptable outflow given the bottleneck capacity, such that the on-set of congestion is prevented;

- Storage space at the ramps and the intersection buffers is fully available for the coordination, meaning that their storage space will only decrease due to outflow reductions resulting from the coordination;

- The ramp itself should not become the bottleneck when buffers are releasing their vehicles after they ran out of storage space. To this aim, the total outflow of the intersection buffers to the ramp is reduced in case it exceeds the ramp's flush metering rate.

4.3 Defining the controlled outflow of storage spaces

By decreasing the inflow to a bottleneck, undesired traffic flow phenomena can be prevented or postponed from happening. To realize such required inflow reduction or *metering task on the bottleneck*, vehicles that travel towards the bottleneck should be held back at storage spaces upstream of the bottleneck by means of coordinated or integrated traffic control. A *metering task on the bottleneck* can also be seen as the buffering request (in terms of vehicles per hour) posed to upstream located controllers, that are able to reduce the flows downstream. In case multiple buffers are used, the outflows of the buffers need to be determined such that:

- The total required inflow reduction to the bottleneck is constantly realized over time;

- All available storage space becomes fully utilized to maximize the metering duration on the bottleneck.

Before discussing how to identify the outflow distribution over the controlled storage spaces, first the overall metering task is specified that keeps a freeway bottleneck at capacity or a ramp queue at constant length.

4.3.1 Keeping a freeway bottleneck at capacity

In case we apply local or coordinated ramp metering (rm) to prevent a flow breakdown at a freeway bottleneck, the metering task $q_{\text{rm}}^{\text{task}}$ is defined by the difference between the bottleneck flow capacity and the total demand towards the bottleneck coming from the downstream freeway stretch and downstream located ramps

$$q_{\text{rm}}^{\text{task}} = \phi_0^{\text{p}}\gamma_0 + \sum_{r\in R}\sum_{b\in B_r}\phi_b^{\text{p}}\alpha_b\gamma_b - c_{\text{bn}}, \tag{4.1}$$

with ϕ_0^{p} and ϕ_b^{p} the flows entering the network from the origins located at the freeway and the intersection buffers $b \in B_r$ during the peak period $0 < t < t^{\text{p}}$, γ_0 the fraction of traffic from the freeway origin traveling past the bottleneck, α_b the fraction of traffic moving from buffer b to the ramp r, and γ_b the fraction of *buffer outflow to the ramp* that moves past the freeway bottleneck, and c_{bn} the freeway bottleneck capacity (see also Figure 4.3). If the metering task on the bottleneck is realized by multiple coordinated ramps, the metering task per ramp q_r^{task} needs to be found such that all storage space becomes efficiently used. The corresponding approach and the way to determine the actual ramp outflow of each coordinated ramp is discussed in Section 4.3.3.

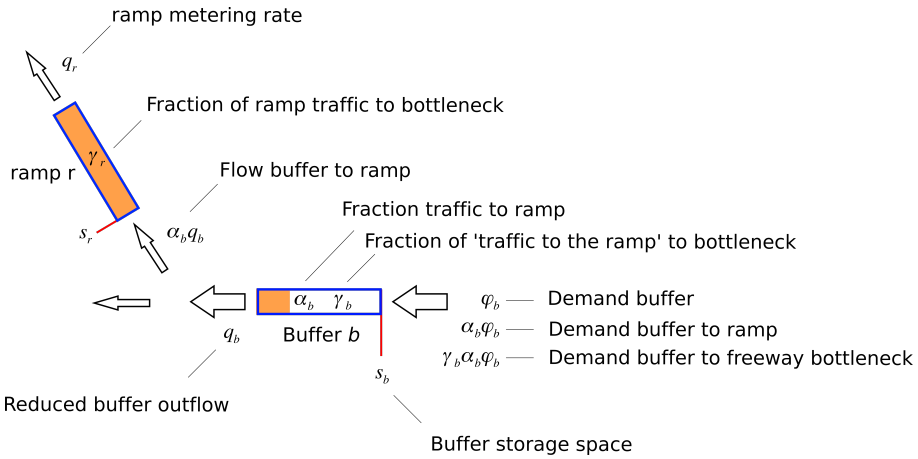

Figure 4.3: Overview of the variables related to the storage space characteristics such as destination dependent fractions, demands and the actual buffer outflows.

4.3.2 Keeping the ramp queue at constant length

In this section the collective metering task of the intersection buffers $q_{\mathrm{ic},r}^{\mathrm{task}}$ upstream of ramp r is discussed to keep a ramp queue at constant length during the peak period by means of integrated ramp metering and intersection control. To this aim, the ramp inflow needs to become equal to the ramp outflow or ramp metering rate (see also Figures 4.2). The outflow reduction that needs to be realized by all coordinated buffers at the intersection (ic) is then given by

$$q_{\mathrm{ic},r}^{\mathrm{task}} = \sum_{b \in B_r} \phi_b^{\mathrm{p}} \alpha_b - q_r, \tag{4.2}$$

with $\sum_{b \in B_r} \phi_b^{\mathrm{p}} \alpha_b$ the peak demand coming from the buffers $b \in B_r$ towards the ramp and q_r the ramp metering rate. To effectively use all storage space, the coordination algorithm should subsequently determine the contribution of each individual intersection buffer $q_{\mathrm{b}}^{\mathrm{task}}$ to this overall task.

4.3.3 Effective utilization of coordinated storage spaces

This section discusses the generalized steps on how to effectively distribute an overall metering task on a bottleneck over multiple controlled storage spaces. The procedure helps us to determine if the problem at hand is solvable, and if so, what the outflows of the buffers become to fully use their available space. The *bottleneck* in this procedure indicates either a *freeway bottleneck* in case of coordinated ramp metering or a *saturated ramp* in case of integrated ramp metering and intersection control. The coordinated and uncoordinated network elements x that feed traffic to the considered bottleneck are gathered in set X. Hence, for coordinated ramp metering, set X consists of on-ramps $r \in R$, and for integrated ramp metering and intersection control it consists of intersection buffers $b \in B_r$ located upstream of ramp r.

To find feasible metering tasks (if possible), it is important to consider the maximum task a buffer can realize. For instance, a task can become infeasible if a buffer has a relatively large amount of storage space available, but a low demand to the bottleneck. The required outflow reduction to fully utilize all space might then become larger than the demand itself. More specifically, the maximum task a buffer can realize, is determined by its

demand ϕ_x^{p}, its minimum outflow rate $q_x^{\min}$ and the share of traffic y_x moving towards the considered bottleneck (i.e. y_x is equal to γ_r for coordinated ramp metering and α_b for integrated ramp metering and intersection control)

$$q_x^{\mathrm{maxtask}} = (\phi_x^{\mathrm{p}} - q_x^{\min}) y_x. \tag{4.3}$$

The part of the metering task that cannot be realized then needs to be redistributed over the other buffers, which in turn might result in the situation where newly given tasks cannot be realized by the other coordinated buffers. A feasible distribution of individual metering tasks and corresponding actual outflows can be found by solving the following procedure in an iterative way.

Step 1: Initiation

All buffers that are adopted in the coordination are gathered in the set $X^{\mathrm{c}} \subseteq X$. Moreover, during the procedure set $A^1 \subseteq X^{\mathrm{c}}$ contains all coordinated buffers that are able to realize their assigned metering task, and set $A^0 \subseteq X^{\mathrm{c}}$ contains all buffers that are not able to realize their given metering task. Initially $A^1 = X^{\mathrm{c}}$ and $A^0 = \varnothing$, but at the end of the procedure they are redetermined.

Step 2: Definition of the metering duration

The metering task is distributed based on the saturation times of the buffers, i.e. the duration it takes for the buffers to become filled given their assigned metering task. The saturation times T_x in terms of hours are therefore calculated for all $x \in X^{\mathrm{c}}$ in case each buffer would be individually responsible for realizing the overall metering task on the bottleneck

$$T_x = \begin{cases} \dfrac{s_x}{(q_{\mathrm{tot}}^{\mathrm{task}} - \sum\limits_{x \in A^0} q_x^{\mathrm{maxtask}})/y_x} & \text{if } x \in A^1 \\ \dfrac{s_x}{q_x^{\mathrm{maxtask}}/y_x} & \text{if } x \in A^0, \end{cases} \tag{4.4}$$

with s_x the storage space at buffer x in terms of number of vehicles, $q_{\mathrm{tot}}^{\mathrm{task}}$ the overall metering task and the denominator terms describing the *effective outflow reduction* buffers need to realize. The summation of the saturation times over all buffers that are able to realize their given metering task, indicates the total duration T^{c} with control can be applied

$$T^{\mathrm{c}} = \sum_{x \in A^1} T_x. \tag{4.5}$$

Buffers that are not able to realize their task influence the total control duration by reducing the remaining task for the buffers that can.

Step 3: Definition of metering tasks

The individual metering task for all buffers is then given by:

$$q_x^{\mathrm{task}} = \begin{cases} \frac{T_x}{T^{\mathrm{c}}}(q_{\mathrm{tot}}^{\mathrm{task}} - \sum_{x \in A^0} q_x^{\mathrm{maxtask}}) & \text{if } x \in A^1 \\ q_x^{\mathrm{maxtask}} & \text{if } x \in A^0 \\ 0 & \text{if } x \notin X^{\mathrm{c}}. \end{cases} \tag{4.6}$$

All coordinated buffers that are able to realize their task are filled at the same moment, and the ones that are not able to do so will not become completely filled.

Step 4: Updating buffer sets

In case a buffer's metering task is larger than its maximum task, sets A^1 and A^0 are updated by

$$A^0 = \{x | q_x^{\mathrm{task}} \geq q_x^{\mathrm{maxtask}}, \forall x \in X^{\mathrm{c}}\} \tag{4.7}$$

and

$$A^1 = X^{\mathrm{c}} - A^0. \tag{4.8}$$

If these sets change with respect to their previous content, the above procedure is repeated starting at step 1, otherwise the overall metering task is properly distributed. If A^1 is empty at the end of the procedure, the metering task cannot be realized given the prevailing situation.

4.3.4 The actual outflows of controlled storage locations

Once the metering tasks are known, the actual outflows or metering rates q_x of the buffers $x \in X$ can be determined by

$$q_x = \phi_x^{\mathrm{p}} - \frac{q_x^{\mathrm{task}}}{y_x}. \tag{4.9}$$

As can be seen in 4.9 the outflows are a function of the buffer's inflow, its task and the fraction of traffic moving to the considered bottleneck. Note that with respect to realizing a certain metering task – a lower fraction of traffic to the bottleneck, requires a lower actual metering rate or buffer outflow. In other words, more vehicles need to be held back that do not travel towards the bottleneck to realize the required metering task at the bottleneck location.

4.4 Definition of the cumulative curves

At this stage, all information is available to build the cumulative curves that enable us to find the set of buffers that minimizes the total system delay. Network, control and demand characteristics such as the buffer sizes, fractions of traffic that move to the freeway, the peak and off-peak demands and typical outflows due to the coordination process are assumed known. Note that in the Section 4.3, it has been discussed how the metering rate (i.e. the controlled outflow) of coordinated buffers is determined.

The design of the curves is shown in Figure 4.4, describing the inflow and outflow over time at the locations where delays can occur (i.e. intersection buffers, on-ramp(s) and freeway bottleneck). The typical flow phases corresponding to the coordination process are identified by the lines with constant slope. Moreover, in the explanation we distinguish the cases ***A*** *Coordinated ramp metering* and ***B*** *Integrated ramp metering and intersection control*.

Intersection buffer inflows

A and **B**: For both cases the intersection buffer inflow curves are similar. Traffic enters the intersection buffers to subsequently flow to the ramp or other urban destinations. As can be seen in Figure 4.4a and d, the inflow curve for an intersection buffer is composed out of lines a_b and b_b, respectively describing the cumulative inflow during the following phases:

- **Peak period (a_b):** The inflow into the buffers, hence the slope of line a_b is determined by a buffer-specific peak period (p) demand ϕ_b^{p} in terms of vehicles per hour;

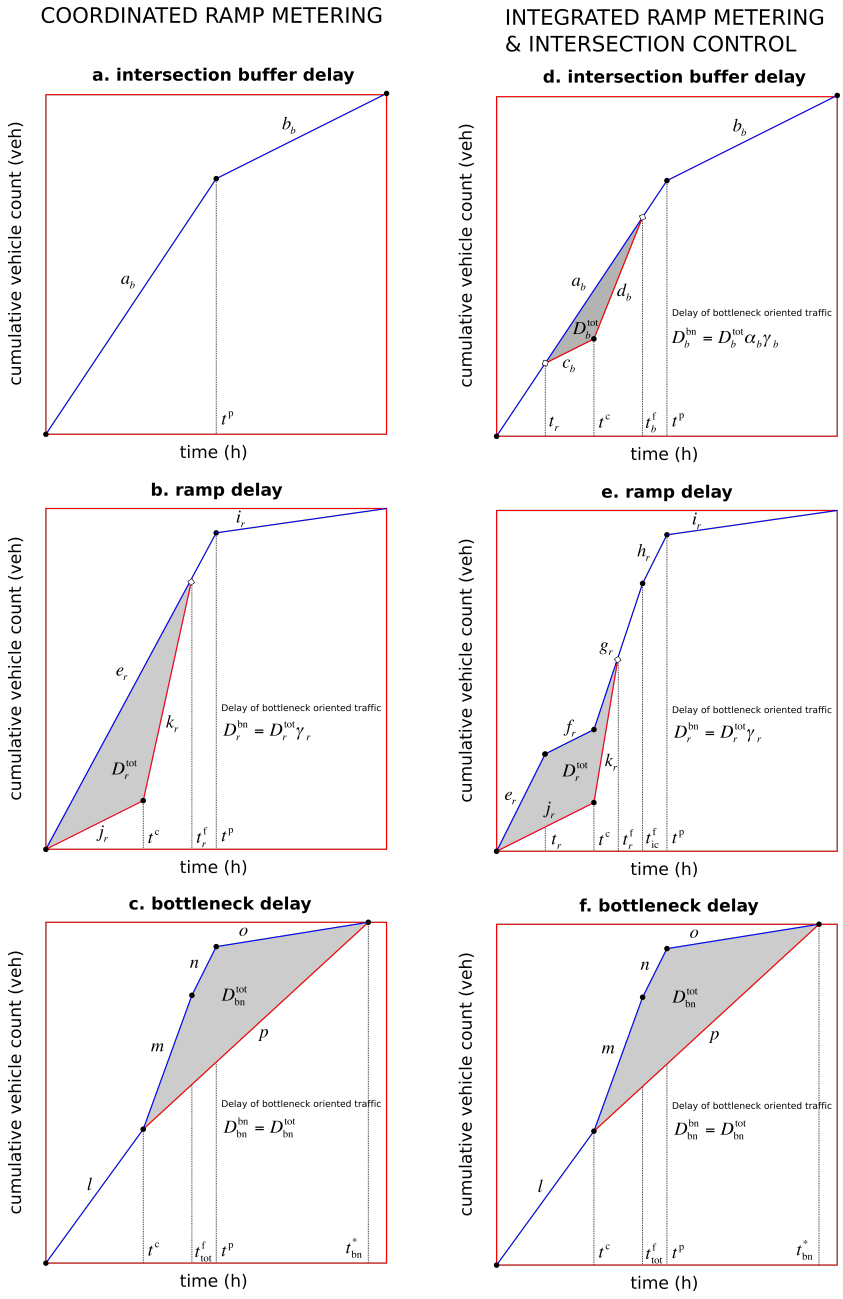

Figure 4.4: The inflow curves in blue and outflow curves in red at key locations in the network for both storage space selection applications with (a,d) the intersection buffers, (b,e) the ramps, and (c,f) the freeway bottleneck.

- **Off-peak period (b_b):** At time $t = t^{\mathrm{p}}$ the peak period demand changes into the off-peak period (o) demand ϕ_b^{o} representing the slope of line b_b.

The buffer inflow over time is then assumed to be given by

$$q_b^{\mathrm{in}}(t) = \begin{cases} \phi_b^{\mathrm{p}} & \text{if } 0 < t \leq t^{\mathrm{p}} \\ \phi_b^{\mathrm{o}} & \text{if } t > t^{\mathrm{p}}. \end{cases} \tag{4.10}$$

Intersection buffer outflows

A: When we are choosing ramps within a coordinated ramp metering scheme (see Figure 4.4a), than the outflow of the intersection buffers during the peak and off-peak period is assumed to be equal to their inflow, because the outflow of the intersection buffers is not reduced by means of control to prevent on-ramp saturation

$$q_b^{\mathrm{out}}(t) = q_b^{\mathrm{in}}(t). \tag{4.11}$$

B: In case we are choosing intersection buffers to prevent further spill-back of a ramp queue (see Figure 4.4d), the buffer outflow curve typically deviates from the inflow curve by the lines c_b and d_b, respectively indicating the following phases:

- **Coordination period (c_b):** When the ramp is saturated at time $t_r = T_r$, the coordination is activated until $t^{\mathrm{c}} = T_r + T^{\mathrm{c}}$, the moment that the ramp and all coordinated intersection buffers are filled. Equations (4.4) to (4.9) describe how the outflow q_b of coordinated and uncoordinated buffers $b \in B_r$ is determined, indicating the slopes of lines c_b.

- **Flush period buffer (d_b):** Vehicles that have been stored during the coordination period are released by increasing the outflow of coordinated intersection buffers to a buffer specific maximum rate q_b^{f} (i.e. flush metering rate) until the queue is dissolved at time t_b^{f}. This rate indicates the slope of line d_b and for uncoordinated buffers it is determined by the buffer demand.

The buffer outflow over time then becomes:

$$q_b^{\text{out}}(t) = \begin{cases} q_b^{\text{in}}(t) & \text{if } 0 < t \leq t_r \\ q_b & \text{if } t_r < t \leq t^{\text{c}} \\ q_b^{\text{f}} & \text{if } t^{\text{c}} < t \leq t_b^{\text{f}} \\ q_b^{\text{in}}(t) & \text{if } t > t_b^{\text{f}}. \end{cases} \tag{4.12}$$

Ramp inflow

A: Ramp inflow is defined by the traffic flows coming from all upstream located intersection buffers $b \in B_r$ that turn to the ramp r. For the coordinated ramp metering case the inflow curve simply consists of lines e_r and i_r describing the traffic demand from the urban network to the ramp, i.e the case that no traffic is held back at the intersection buffers.

- **Peak period (e_r):** The inflow during the peak period is defined by the peak demand coming from all involved intersection arms $b \in B_r$ that turns towards the freeway during the peak period, i.e. $\phi_b^{\text{p}}\alpha_b$;
- **Off-peak period (i_r):** The inflow during the off-peak period for $t > t^{\text{p}}$ is subsequently determined by the off-peak demand.

The ramp inflow in this case becomes

$$q_r^{\text{in}}(t) = \begin{cases} \sum\limits_{b \in B_r} \phi_b^{\text{p}}\alpha_b & \text{if } 0 < t \leq t_{\text{p}} \\ \sum\limits_{b \in B_r} \phi_b^{\text{o}}\alpha_b & \text{if } t > t^{\text{p}}. \end{cases} \tag{4.13}$$

B: When dealing with the integrated ramp metering and intersection control case, the curve is more complex, because its inflow coming from upstream located buffers is controlled to keep the ramp queue at constant length for as long as possible. As can be seen in Figure 4.4e the cumulative inflow curve consists of the additional lines f_r, g_r, h_r, representing the following inflow phases:

- **Coordination period (f_r):** At the moment the ramp becomes saturated at $t = t_r$, the intersection buffers start limiting the inflow to the ramp. To this aim, the ramp's inflow is synchronized with the ramp's outflow (determined by the ramp metering rate q_r) until all intersection buffers have run out of space at time t^{c};

- **Flush period buffers (g_r):** When all storage space is used, coordinated intersection buffers release their stored vehicles. However, their queues might not be dissolved at the exact same time. This might be due to varying storage space size or maximum outflow rate. A flush period is therefore considered that describes the period between the end of the coordination and the moment that all buffer queues at the intersection are dissolved, i.e. $t^{\mathrm{c}} < t \leq t^{\mathrm{f}}_{\mathrm{ic},r}$ where $t^{\mathrm{f}}_{\mathrm{ic},r}$ is defined by the moment that the last buffer queue is dissolved $t^{\mathrm{f}}_{\mathrm{ic},r} = \max_{b \in B^{\mathrm{c}}_r} t^{\mathrm{f}}_b$. The corresponding average flow into the ramp during this period is indicated by $q^{\mathrm{f}}_{\mathrm{ic},r}$ and discussed in the next paragraph;

- **Remaining peak period (h_r):** In case the buffer queues are dissolved before the end of the peak period, the inflow into the ramp for $t^{\mathrm{f}}_{\mathrm{ic},r} < t \leq t^{\mathrm{p}}$ will be determined by the peak period demand.

The flow $q^{\mathrm{f}}_{\mathrm{ic},r}$ that enters the ramp during the flush period $t^{\mathrm{c}} < t \leq t^{\mathrm{f}}_{\mathrm{ic},r}$ is determined by means of the cumulative outflow curves of the buffers

$$q^{\mathrm{f}}_{\mathrm{ic},r} = \frac{\sum_{b \in B_r} N^{\mathrm{f}}_b \alpha_b}{t^{\mathrm{f}}_{\mathrm{ic},r} - t^{\mathrm{c}}}, \tag{4.14}$$

with N^{f}_b the number of vehicles leaving buffer $b \in B_r$ (i.e. coordinated and uncoordinated ones). Note that the fourth period (remaining peak period after coordination) described by line h_r, will not be defined when the flush period of the intersection buffers ends in the off-peak period, i.e. when $t^{\mathrm{f}}_{\mathrm{ic},r} > t^{\mathrm{p}}$. In that case, the ramp inflow for $t > t^{\mathrm{f}}_{\mathrm{ic},r}$ will be defined by the off-peak period demand. The ramp inflow curve can then be written as:

$$q^{\mathrm{in}}_r(t) = \begin{cases} \sum_{b \in B_r} \phi^{\mathrm{p}}_b \alpha_b & \text{if } 0 < t \leq t_r \\ \sum_{b \in B_r} q_b \alpha_b & \text{if } t_r < t \leq t^{\mathrm{c}} \\ q^{\mathrm{f}}_{\mathrm{ic},r} & \text{if } t^{\mathrm{c}} < t \leq t^{\mathrm{f}}_{\mathrm{ic},r} \\ \sum_{b \in B_r} \phi^{\mathrm{p}}_b \alpha_b & \text{if } t^{\mathrm{f}}_{\mathrm{ic},r} < t \leq t^{\mathrm{p}} \\ \sum_{b \in B_r} \phi^{\mathrm{o}}_b \alpha_b & \text{if } t > t^{\mathrm{p}}. \end{cases} \tag{4.15}$$

Ramp outflow

A and **B**: The ramp outflow curves of both approaches are similar. As can be seen in Figure 4.4b, a ramp's outflow curve consists of lines j_r and k_r, indicating the situation where the ramp is metering on the bottleneck to prevent a flow breakdown with metering rate q_r, and the situation where stored ramp vehicles are released into the mainstream with flush metering rate q_r^{f} to prevent further spill-back of the ramp queue or because the conditions at the freeway become undersaturated.

- **Metering period ramp (j_r)**: As long as there is storage space available at either the ramp or the intersection buffers, the ramp will be metering on the freeway bottleneck. Its outflow until the end of the coordination at t^{c} is therefore equal to the required metering rate q_r to keep the freeway bottleneck at capacity.
- **Flush period ramp (k_r):** In case storage space has run out, the ramp outflow is increased to a predefined flush metering rate q_r^{f}, causing congestion in the bottleneck. However, when the ramp outflow is increased due to undersaturated conditions, its maximum is constrained by the ramp capacity or the acceptable outflow given the bottleneck capacity (such that the on-set of congestion is prevented). Hence, the metering task on the bottleneck and corresponding outflow is determined by (4.1) for the off-peak demands.

From the moment the ramp queue is dissolved at t_r^{f}, the ramp outflow curve becomes equal to its inflow curve. The ramp outflows are thus given by

$$q_r^{\mathrm{out}}(t) = \begin{cases} q_r & \text{if } 0 < t \leq t^{\mathrm{c}} \\ q_r^{\mathrm{f}} & \text{if } t^{\mathrm{c}} < t \leq t_r^{\mathrm{f}} \\ q_r^{\mathrm{in}}(t) & \text{if } t > t_r^{\mathrm{f}}. \end{cases} \tag{4.16}$$

Bottleneck inflow

A and **B**: The benefits of the coordination appear at the bottleneck, i.e. the longer the freeway capacity drop is prevented, the smaller the total bottleneck delay becomes. Postponing the capacity drop can be done by including more on-ramps or intersection storage spaces into the coordination. In Figure 4.4c, it can be seen that the bottleneck inflow curve consists of lines l, m, n and o, representing the following phases:

- **Local and coordinated metering period (l):** A freeway bottleneck can be kept at capacity as long as there is storage space available to hold back traffic that moves towards the bottleneck for $0 < t \leq t^{\mathrm{c}}$;
- **Flush period storage (m):** The flush period $t_c < t < t^{\mathrm{f}}_{\mathrm{tot}}$ is defined as the duration it takes to release all stored vehicles, i.e. when holding back vehicles by means of coordinated ramp metering $t^{\mathrm{f}}_{\mathrm{tot}} = \max_{r \in R^{\mathrm{c}}}(t^{\mathrm{f}}_r)$, and when holding back vehicles at a ramp's upstream located intersection $t^{\mathrm{f}}_{\mathrm{tot}} = \max(t^{\mathrm{f}}_r, t^{\mathrm{f}}_{\mathrm{ic},r})$. The corresponding total flow into the bottleneck $q^{\mathrm{f}}_{\mathrm{tot}}$ is discussed in the next paragraph;
- **Remaining peak period (n):** If the ramp's total flush period ends within the peak period $t^{\mathrm{f}}_{\mathrm{tot}} < t^{\mathrm{p}}$, then peak period demand flows into the bottleneck for $t^{\mathrm{f}}_{\mathrm{tot}} < t \leq t^{\mathrm{p}}$;
- **Off-peak period (o):** The bottleneck inflow during the off-peak period for $t > t^{\mathrm{p}}$ is determined by the off-peak demand.

The inflow of the bottleneck $q^{\mathrm{f}}_{\mathrm{tot}}$ during the total flush period $t^c < t < t^{\mathrm{f}}$ is defined based on the parameterized outflow curves $f^{\mathrm{out}}_r(t)$ of involved ramps r and the inflow at the freeway origin. The number of vehicles N^{f}_r that flows out of involved ramps $r \in R$ is then defined by

$$N^{\mathrm{f}}_r = f^{\mathrm{out}}_r(t^{\mathrm{f}}_{\mathrm{tot}}) - f^{\mathrm{out}}_r(t^{\mathrm{c}}), \tag{4.17}$$

and the number of vehicles that enters the freeway at its downstream origin becomes

$$N^{\mathrm{f}}_0 = \begin{cases} \phi^p_0(t^{\mathrm{f}}_{\mathrm{tot}} - t^{\mathrm{c}}) & \text{if } t^{\mathrm{f}}_{\mathrm{tot}} \leq t^{\mathrm{p}} \\ \phi^p_0(t^{\mathrm{p}} - t^{\mathrm{c}}) + \phi^o_0(t^{\mathrm{f}}_{\mathrm{tot}} - t^{\mathrm{p}}) & \text{if } t^{\mathrm{f}}_{\mathrm{tot}} > t^{\mathrm{p}}. \end{cases} \tag{4.18}$$

The average flow that arrives at the bottleneck during the total flush period $q^{\mathrm{f}}_{\mathrm{tot}}$ is then determined by line m with slope

$$q^{\mathrm{f}}_{\mathrm{tot}} = \frac{\sum_{r \in R} N^{\mathrm{f}}_{\mathrm{r}} \gamma_r + N^{\mathrm{f}}_0 \gamma_0}{t^{\mathrm{f}}_{\mathrm{tot}} - t^{\mathrm{c}}}, \tag{4.19}$$

with N^{f}_r the number of vehicles leaving the ramp and N^{f}_0 the number of vehicles coming from the freeway origin for $t^{\mathrm{c}} < t \leq t^{\mathrm{f}}_{\mathrm{tot}}$. Note that the third period (remaining peak period after coordination) described by line n, will

not be defined when the flush period of the ramp ends in the off-peak period, i.e. when $t^{\mathrm{f}}_{\mathrm{tot}} > t^{\mathrm{p}}$. In that case, the inflow for $t > t^{\mathrm{f}}_{\mathrm{tot}}$ is determined by the off-peak demand. The bottleneck inflow curve can then be written as

$$q^{\mathrm{in}}_{\mathrm{bn}}(t) = \begin{cases} \gamma_0 \phi^{\mathrm{p}}_0 + \sum\limits_{r \in R} \gamma_r q_r & \text{if } 0 < t \leq t^{\mathrm{c}} \\ q^{\mathrm{f}}_{\mathrm{tot}} & \text{if } t^{\mathrm{c}} < t \leq t^{\mathrm{f}}_{\mathrm{tot}} \\ \phi^{\mathrm{p}}_0 \gamma_0 + \sum\limits_{r \in R} \sum\limits_{b \in B_r} \phi^{\mathrm{p}}_b \alpha_b \gamma_b & \text{if } t^{\mathrm{f}}_{\mathrm{tot}} < t \leq t^{\mathrm{p}} \\ \phi^{\mathrm{o}}_0 \gamma_0 + \sum\limits_{r \in R} \sum\limits_{b \in B_r} \phi^{\mathrm{o}}_b \alpha_b \gamma_b & \text{if } t > t^{\mathrm{p}}. \end{cases} \tag{4.20}$$

Bottleneck outflow

A and **B**: In Figure 4.4c, the bottleneck outflow curve is described by lines l and p, respectively representing the phases:

- **Local and coordinated metering period (l):** As long as the bottleneck is kept at capacity by the ramp meter for $0 < t \leq t^{\mathrm{c}}$, its outflow is equal to the inflow;
- **Flow breakdown period (p):** After the coordination for $t > t^{\mathrm{c}}$, the bottleneck outflow drops with a certain percentage f^{cdrop} due to the capacity drop phenomenon.

The bottleneck outflow curve is then described by

$$q^{\mathrm{out}}_{\mathrm{bn}}(t) = \begin{cases} c_{\mathrm{bn}} & \text{if } t \leq t^{\mathrm{c}} \\ c_{\mathrm{bn}} f^{\mathrm{cdrop}} & \text{if } t > t^{\mathrm{c}}. \end{cases} \tag{4.21}$$

A fixed freeway capacity drop is assumed, independent of the severity of the congestion (i.e. independent of the outflow of the ramp). Under such conditions it is important to reduce the hindrance to vehicles at the ramp and urban network as fast as possible from the moment that storage space has run out and the freeway capacity has dropped. Note that a queue protection mechanism that sets the ramp and buffer outflow equal to the inflow as in Spiliopoulou et al. (2010), would in this case cause unnecessary delays at the buffers and is therefore not incorporated.

4.5 Quantification of delays

All variables that influence intersection buffer, ramp and bottleneck delay are now related by means of the vectorized graphs, such that their relation to the total system delay can be evaluated. The required steps are elaborated in this section, while using the intersection buffer delay to illustrate the computational steps.

First the specific lines in the inflow and outflow curve are defined that identify the point of intersection and the delay area. For instance, in Figure 4.4a it can be seen that point $(t_b^{\mathrm{f}}, n_b^{\mathrm{f}})$ defines the area between the intersection buffer curves, i.e. where line d_b crosses either line a_b or b_b; indicating the moment t_b^{f} where the buffer queue is dissolved and the buffers' flush period comes to an end.

The point of intersection between the considered lines is determined by parameterizing them and subsequently solving the system of equations in the form $A\vec{x} = \vec{z}$ by using an inverse multiplication $\vec{x} = A^{-1}\vec{z}$. In our worked example, the line equation $f_{a_b}(t)$ representing the cumulative vehicle count n over time t for intersection buffer b can be written as

$$n - a_b t = 0, \tag{4.22}$$

the line equation $f_{b_b}(t)$ with $n = a_b t^{\mathrm{p}} + b_b(t - t^{\mathrm{p}})$ as

$$n - b_b t = (a_b - b_b)t^{\mathrm{p}}, \tag{4.23}$$

and the line equation $f_{d_b}(t)$ with $n = c_b t_c + d_b(t - t^{\mathrm{c}})$ as

$$y - d_b t = (c_b - d_b)t^{\mathrm{c}}. \tag{4.24}$$

The system of equations for determining the point of intersection between lines a_b and d_b becomes

$$\begin{bmatrix} -a_b & 1 \\ -d_b & 1 \end{bmatrix} \begin{bmatrix} t_b^{\mathrm{f}} \\ n_b^{\mathrm{f}} \end{bmatrix} = \begin{bmatrix} 0 \\ (c_b - d_b)t^{\mathrm{c}} \end{bmatrix}. \tag{4.25}$$

As can be seen in Figure 4.4a, the point of intersection lies before t_{p}, meaning that we are dealing with a triangular shaped delay area and that our point of interest is found. This enables us to determine the buffer's delay by means

of the cross product between the involved lines in vector format (e.g. $\vec{a}_b = [1, a_b]'$) that enclose the area

$$D_b^{\text{tot}} = \frac{||\vec{a}_b t_b^{\text{f}} \times \vec{c}_b t^{\text{c}}||}{2}. \tag{4.26}$$

Notice that if the point of intersection between lines a_b and d_b lies beyond t^{p}, the resulting area would become quadrangle. Then, the point of intersection between lines b_b and d_b needs to be determined by solving the following system of equations

$$\begin{bmatrix} -b_b & 1 \\ -d_b & 1 \end{bmatrix} \begin{bmatrix} t_b^{\text{f}} \\ n_b^{\text{f}} \end{bmatrix} = \begin{bmatrix} (a_b - b_b)t^{\text{p}} \\ (c_b - d_b)t^{\text{c}} \end{bmatrix}. \tag{4.27}$$

The resulting total delay at intersection buffer b would then be given by

$$D_b^{\text{tot}} = \frac{||\vec{a}_b t^{\text{p}} \times \vec{c}_b t^{\text{c}}||}{2} + \frac{||\vec{d}_b (t_b^{\text{f}} - t^{\text{c}}) \times \vec{b}_b (t_b^{\text{f}} - t^{\text{p}})||}{2}. \tag{4.28}$$

To conclude, the delay caused to vehicles that travel along the bottleneck is given by

$$D_b^{\text{bn}} = \alpha_b \gamma_b D_b^{\text{tot}}, \tag{4.29}$$

and the delay to vehicles that do not travel along the bottleneck is given by

$$D_r^{\text{elsewhere}} = (1 - \alpha_b \gamma_r) D_b^{\text{tot}}. \tag{4.30}$$

The ramp and bottleneck delay are derived in a similar way. The total system delay is then defined by the sum of the delays caused at the ramps, the intersection buffers and the freeway bottleneck

$$D_{\text{sys}}^{\text{tot}} = \sum_{r \in R^{\text{c}}} \sum_{b \in B_r^{\text{c}}} D_b^{\text{tot}} + D_r^{\text{tot}} + D_{\text{bn}}^{\text{tot}}. \tag{4.31}$$

All variables that influence the total ramp and bottleneck delay are now related by means of cumulative curves, such that their relation to the total system delay can be evaluated. In the remainder of this section we will discuss the size of bottleneck delay at the freeway and the potential impact of ramp metering on it.

The bottleneck delay $D_{\text{bn}}^{\text{tot}}$ in case no control is applied (i.e. the potential bottleneck delay), is predominantly determined by the duration of the peak period and the size of the capacity drop, as can be seen in Figure 4.5a. The

longer the oversaturated peak period t^{p} lasts, the larger the bottleneck delay becomes (arrow 1 and red area). The same holds for the size of the capacity drop that negatively influences the slope of the outflow curve (arrow 2 and yellow area). Hence, bottleneck delay grows more than linear as a function of increasing peak period and capacity drop.

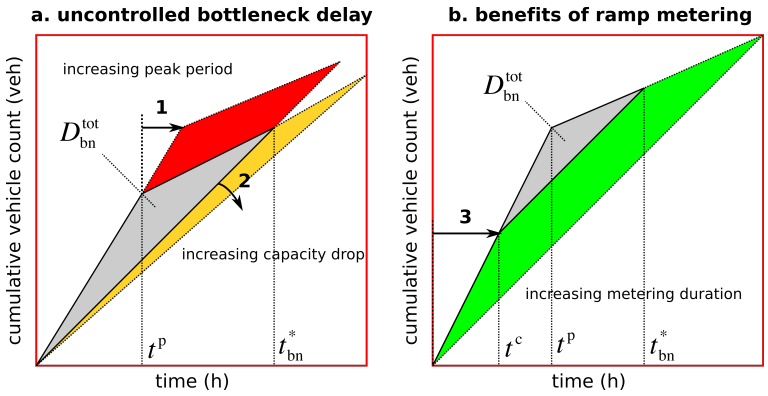

Figure 4.5: (a) Impact of increasing peak period duration and size of the capacity drop on the bottleneck delay for the uncontrolled case, and (b) benefits of preventing the capacity drop by means of ramp metering.

By making more buffer space available, the metering duration on the bottleneck is extended reducing its delay, as can be seen in Figure 4.5b (arrow 3 and green area). The bottleneck delay decreases fast in the beginning of the metering period, and the longer the metering period becomes, the smaller the gains per extra metered time unit. This is interesting, because this implies that the benefits of adding extra buffer space to extend the metering duration on the bottleneck becomes less, while the caused delays to vehicles that do not pass the bottleneck become higher for each extra buffer added to the coordination (i.e. considered in the order of decreasing traffic fraction to the bottleneck).

4.6 Test case: Ramp storage space selection

By means of a simulation test case it is shown how the proposed approach can be used to determine the optimal set of ramps for a coordinated ramp metering scheme given the prevailing conditions. Moreover, the impact of

various network and demand characteristics will be explored to gain a good understanding on how these variables impact the effectiveness of the coordination. To enable the evaluation of different research questions, the modeling scenarios have been parameterized and implemented into the computational model of the approach.

The network layout for coordinated ramp metering case is in line with the situation given in Figure 4.1a. Traffic flows via intersection buffers to the ramps and to the freeway bottleneck. During the peak hour the bottleneck becomes oversaturated, resulting in a capacity dropped network outflow for as long as there is congestion. During the off-peak period, demands are chosen such that the freeway congestion and the coordinated ramp queues will dissolve. The following questions are explored by means of the proposed approach:

- Scenario A: How do the capacity drop and the duration of the peak period impact potential freeway bottleneck delay?
- Scenario B: How does a ramp's traffic fraction to the bottleneck impacts the decision of adopting it into the coordination?
- Scenario C: How does ramp storage space impact the effectiveness of the coordination?
- Scenario D: What is the optimal set of ramps within a coordinated ramp metering scheme given the conditions at hand?

A detailed overview of the parameters used for the different test case scenario's is given in Table 4.1. The parameter values, such as the demands, bottleneck capacity and capacity drop value are chosen such that they make a meaningful scenario with well interpretable results. For instance, the problem needs to be solvable, i.e. ramps need to be able to realize the metering task on the bottleneck for a significant amount of time. In the first scenario *A* we evaluate the potential bottleneck delay when no control is applied. Moreover, to keep the the results comprehensible scenarios *B* and *C* consist of only two ramps. The last scenario *D* is a bit more involved to illustrate how an optimal set of coordinated ramps can be found for a network consisting of 5 ramps with varying characteristics.

Because of our focus on choosing ramps and to keep the scenarios well interpretable, we have aggregated the intersection buffer characteristics into ramp specific demands and traffic fractions to the bottleneck. Moreover, to make proper comparisons within a scenario, the metering task on the freeway bottleneck is kept constant. To this aim, the fraction of freeway traffic moving towards the bottleneck is made a dependent variable and therefore not specified in the input table. In the remainder of this section the aspects are discussed that will be analyzed by means of different test case scenarios.

Table 4.1: Set-up of the test case scenarios.

Param	**Unit**	**Scenario A**	**Scenario B**	**Scenario C**	**Scenario D**
t^{p}	h	[0-3]	.25	.5	[.5, 1]
f^{cdrop}	%	[0-20]	10	10	10
c^{bn}	veh/h	4000	4000	4000	4000
$q_{\mathrm{bn}}^{\mathrm{task}}$	veh/h	400	400	400	400
ϕ_0^{p}	veh/h	4000	4000	4000	2400
ϕ_r^{p}	veh/h	[600, 600]	[600,600]	[600,600]	[600,600,600,600,600]
ϕ_0^{o}	veh/h	2000	2000	2000	1200
ϕ_r^{o}	veh/h	[300,300]	[300,300]	[300,300]	[300,300,300,300,300]
s_r	veh	[50,50]	[50,50]	[0-100,0-100]	[50,100,70,60, 50]
γ_r	-	[1,1]	[1,0-1]	[1,1]	[1,.6,.1,.3,.4]
q_r^{f}	veh/h	[1500,1500]	[1500,1500]	[1200,1200]	[1500,1500,1500,1500,1500]
$r \in R^{\mathrm{c}}$	1/0	[0,0]	[1,0/1]	[1,1]	[0/1,0/1,0/1,0/1,0/1]

4.6.1 Scenario A: Impact of capacity drop and peak period

Let us consider the cumulative inflow and outflow curve of the freeway bottleneck for the uncontrolled case as given in Figure 4.5a. The inflow is determined by the peak and off-peak demand and the outflow by the capacity drop. It can be seen that the peak period duration and the size of the capacity drop have a more than linear effect on the total bottleneck delay. This is interesting, because the larger the total bottleneck delay, the higher the reduction in delay becomes by postponing the capacity drop, hence the more effective the coordination.

To illustrate these relations, the bottleneck delay is evaluated as a function of the capacity drop and peak period duration in case no ramp metering is applied and the metering task on the bottleneck is 400 vehicles per hour. Moreover, to illustrate the effectiveness of the coordination, the change in delay is determined in relation to the metering duration and the peak period.

As can be seen in Figure 4.6a, the bottleneck delay increases more than linearly with an increasing duration of the peak period, and the larger the capacity drop the stronger this impact. Peak periods over one hour are very common, which means that the bottleneck delay can become very high if no control would be applied. This in turn implies that also the benefits of postponing the capacity drop become very high.

Figure 4.6b illustrates the change in delay given a peak period duration and a metering duration of the ramp. For instance, the benefits of ramp metering (reduction in bottleneck delay) during the first metering minutes can become 100 veh.h per minute metering for a peak period of 3 hours. The benefits per extra metering minute decrease with an increasing metering duration.

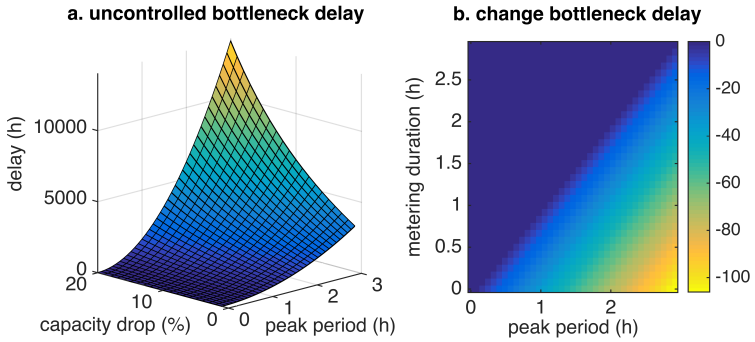

Figure 4.6: (a) Delay at the bottleneck in relation to size of the capacity drop and the duration of the peak period, and (b) change in bottleneck delay as a function of the ramp metering duration and the peak period.

4.6.2 Scenario B: Impact of traffic fraction to the bottleneck

The larger the fraction of traffic with destination bottleneck, the more efficient ramp space is used to reduce the delay at the bottleneck. In other words, if this fraction decreases, more vehicles are delayed that do not have an effect on the bottleneck, and the shorter the coordination will be active. The question is to what extent it is useful to add ramps with such low fraction to the coordination. The resulting ramp, bottleneck and system delay in the coordination case are therefore compared with the situation in which only local control is applied. In this way, the break-even fraction will become apparent that indicates to what extent it is useful to include the second ramp into the coordination. The considered peak period is, however, chosen relatively short, to keep the bottleneck and ramp delays within the same order of magnitude, i.e. to have a break-even fraction in the first place and to clearly show the delays in a single plot. Moreover, this break-even fraction is also determined as a function of the peak period and available ramp space.

In Figure 4.7a and b the ramp, bottleneck and system delay are given for the coordinated and local ramp metering situation as a function of the fraction of vehicles from the assisting ramp to the bottleneck. The situation is analysed for respectively 40 and 60 vehicles of storage space at the assisting ramp. In both graphs the following general trends can be seen for the *coordinated* metering situation:

- The total ramp delay (blue solid line) is increasing with an increasing fraction. This is due to the fact that the higher the fraction becomes, the longer the coordination can be active, and the higher the total ramp delay becomes;

- The lower the fraction, the larger the ramp delay caused to vehicles that do not travel along the bottleneck (difference between blue solid and green solid line), the faster the ramp storage space runs out. This in turn shortens the metering duration, causing a higher bottleneck delay (red solid line);

- The bottleneck delay (red solid line) decreases faster than the total ramp delay (blue solid line) increases, hence, the total system delay (purple solid line) decreases with an increasing fraction.

Hence, the question then becomes to what extent it is beneficial to include ramps with low fractions in the coordination. In other words, given a certain fraction, is the delay at an assisting ramp (difference between blue solid and blue dotted line) smaller than the bottleneck delay reduction that is realized by including the ramp into the coordination (difference between red dotted and red solid line).

From the location where the total system delays for the coordinated and local ramp metering situation intersect (purple solid and purple dotted lines), the break-even points can be determined, i.e. for Figure 4.7a and b these are respectively 20 and 22%. This means that if the fraction becomes lower than that, it is more beneficial to take the assisting ramp out of the coordination.

Introducing more ramp space at the assisting ramp (Figure 4.7a versus b) does not influence the delays in the local ramp metering situation (dotted lines), but does influence the coordinated situation. The following can be seen:

- For very low fractions, the total ramp delay (blue solid line) increases more when adding more space to the ramp than the bottleneck delay (red solid line) reduces. Note that the extra metering duration on the bottleneck (and thus the extra delay reduction) is very limited, due to the fact that the storage space is filling at a very high rate for low fractions;
- For higher fractions, the total ramp delay will also increase when adding more ramp space at the assisting ramp. However, due to the high fraction of traffic to the bottleneck, the ramp's space is effectively used to extend the metering duration on the bottleneck. This leads to a higher reduction in bottleneck delay (red solid line) with respect to the local metering situation (red dotted line);
- Introducing more ramp space in this respect leads to a higher minimum (break-even) fraction for the ramp to be beneficial in the coordinated situation.

In Figure 4.7c the break-even fractions are determined as a function of the peak period duration for different storage space size at the assisting ramp. break-even fractions decrease with longer peak periods, meaning that also

ramps with a low fraction might become beneficially included in the coordination. This is due to the fact that the reduction in bottleneck delay per metering unit of time is higher for longer peak periods when postponing the capacity drop.

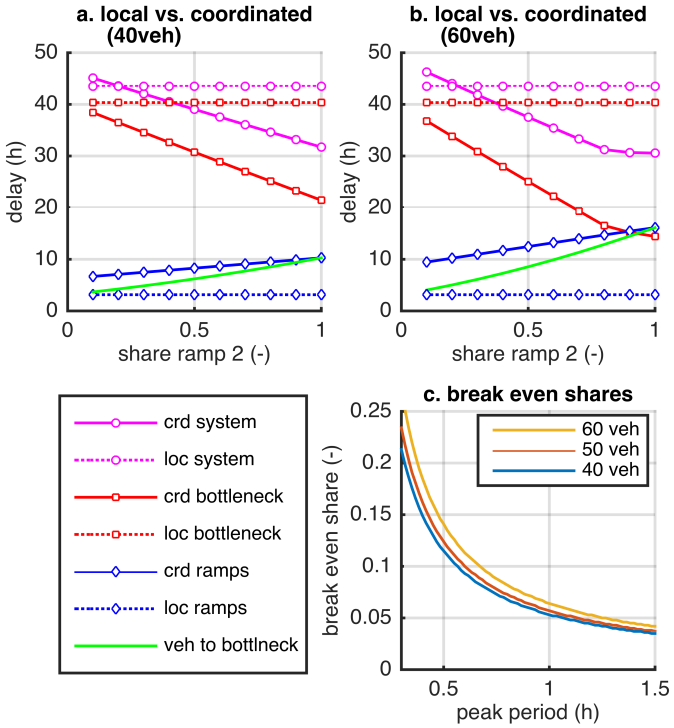

Figure 4.7: (a,b) Ramp, bottleneck and system delay as a function of the ramp traffic fraction to the bottleneck for storage space size at the assisting ramp of respectively 40 veh and 60 veh, and (c) the break-even fractions as a function of the peak period duration for different storage space size of the assisting ramp, indicating the minimum fraction at the assisting ramp that makes coordinating it beneficial.

4.6.3 Scenario C: Impact of ramp storage space

The larger the available space, the longer the bottleneck can be kept at capacity, hence the larger the reduction in bottleneck delay. However, the question is whether storing a large number of vehicles can also negatively influence

the network performance. This is evaluated by determining the ramp, bottleneck and system delay when applying coordination with different combinations of ramp storage space.

In Figure 4.8a, it can be seen that with increasing storage space on both ramps, the ramp delays increase (lower surface), and that the bottleneck delay (upper surface) and thus the total system delay in Figure 4.8b have a minimum. When the peak period comes to an end, and the ramps are still flushing their stored vehicles into the main stream, the bottleneck will remain oversaturated for a longer time than the peak period would essentially cause. In other words, if more traffic is stored than can be flushed into the mainstream between the moment the ramps start flushing and the end of the peak period, the bottleneck delay will increase again. In Figure 4.4d, this happens when $t^{\mathrm{f}}_{\mathrm{tot}} > t^{\mathrm{p}}$ meaning that line g ends later than the beginning of the off-peak period. Hence, the total amount of available storage space can have a negative impact on the total system delay and should therefore be chosen situation-specific.

4.6.4 Scenario D: Optimal ramp configuration

In this scenario it is illustrated that the approach can be used to find an optimal ramp configuration, and that the optimal configuration depends on the duration of the peak period and the size of the capacity drop. To this aim, a 5 ramp network is used, in which each ramp has another storage space size and fraction of traffic moving to the bottleneck. By evaluating all binary ramp configuration combinations, the optimal configuration can be easily found.

In Figure 4.8c and d the optimal set of ramps into the coordination is determined for peak period durations of respectively 0.5 and 1 hour. As can be seen, different sets of coordinated ramps lead to a different total system delay. Moreover, by comparing the optimal solution for both peak periods indicated by the red diamonds, it can be seen that for a longer peak period, it becomes beneficial to adopt ramp 4 in the coordination, which has a relatively low fraction of vehicles towards the bottleneck (see Table 4.3-Sc.D).

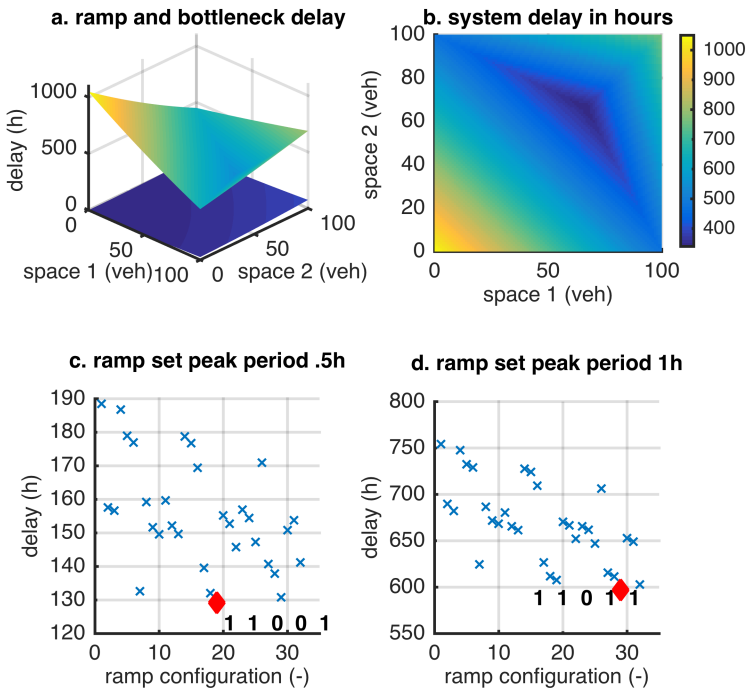

Figure 4.8: (a,b) Ramp, bottleneck and system delay as a function of ramp storage space, and (c,d) System delay given different sets of coordinated ramps for peak periods of 0.5 and 1 hour.

4.6.5 Conclusions

In this section we have presented an approach to determine the optimal ramp configuration when applying coordinated ramp metering. The approach gives clear insight into the factors that influence the fundamental shape and size of the ramp, bottleneck and system delay. At assisting ramps there is a minimum fraction of vehicles towards the bottleneck needed to beneficially adopt it into the coordination scheme. This brake even fraction is determined by the duration of the peak period and the size of the capacity drop. The higher the potential bottleneck delay (in case no ramp metering would be applied), the smaller break-even fractions become. Even ramps with a small fraction of traffic to the bottleneck might then become beneficially adopted in the coordination. It is also shown that storing too many vehicles might negatively influence the total system delay. This can be the case when ramps increase their metering rates to prevent spill-back to the urban roads at the end of the peak period, keeping the bottleneck longer oversaturated (by the large amount of stored traffic flowing into the freeway) than would be the case with respect to the peak period itself.

4.7 Test case: Urban storage space selection

This second simulation test case is used to illustrate how the approach can be used for integrated ramp metering and intersection control given prevailing conditions. The intersection buffers upstream of a ramp are then used to temporarily hold back vehicles that move to the ramp in order to postpone ramp saturation. Moreover, the impact of various network and demand characteristics will be explored to gain a good understanding on how these variables impact the effectiveness of this form of integrated traffic management. To enable the evaluation of different research questions, the scenarios have been parameterized and implemented into the computational model of the approach.

The network layout for the integrated ramp metering and intersection control case is given in Figure 4.1b. Traffic flows via intersection buffers to the ramps and to the freeway bottleneck. During the peak hour the bottleneck becomes oversaturated, resulting in a capacity dropped network outflow for as long as there is congestion. During the off-peak period, demands are cho-

sen such that the freeway congestion and the coordinated ramp queues will dissolve. The following questions are explored by means of the proposed approach:

- Scenario A: How do the capacity drop and the duration of the peak period impact potential freeway bottleneck delay?
- Scenario B: How do intersection buffers' traffic fraction to the bottleneck impact the decision of adopting them into the coordination?
- Scenario C: How does intersection buffers' storage space impact the effectiveness of the coordination?
- Scenario D: What is the optimal set of intersection buffers for the integrated ramp metering and intersection control scheme given the conditions at hand?

In the remainder of this section the aspects are discussed that will be analyzed by means of different test case scenarios in line with the situation description given in Figure 4.2. A detailed overview of the corresponding scenario parameter settings is given in Table 4.2. The parameter values, such as the demands, bottleneck capacity and capacity drop value are chosen such that they make a meaningful scenario with interpretable results. For instance, the problem needs to be solvable, buffers need to be able to realize the metering task on the ramp for a significant amount of time.

4.7.1 Scenario A: Impact of capacity drop, peak period and ramp metering duration

In Scenario A it is shown that the peak period duration and the size of the capacity drop determine the potential bottleneck delay (in case no control would be applied). Moreover, the change in bottleneck delay is shown as a function of the peak period duration and the metering duration, to illustrate that the absolute benefits of ramp metering per metering time unit become smaller with an increasing metering duration. The modeled connection in this scenario consists of a ramp and two feeding upstream intersection arms.

As can be seen in Figure 4.9a, the bottleneck delay increases more than linearly with an increasing duration of the peak period, and the larger the capacity drop the stronger this impact. Peak periods over one hour are very

Table 4.2: Set-up of the test case scenarios.

Param	**Unit**	**Scenario A**	**Scenario B**	**Scenario C**	**Scenario D**
t^{p}	h	0-3	1	1	[1,3]
f^{cdrop}	%	0-20	10	10	10
q_r^{task}	veh/h	400	250	300	300
ϕ_0^{p}	veh/h	3500	5000	4000	4000
ϕ_b^{p}	veh/h	[500,500]	[1000,1000,1000]	[1000, 1000, 1000]	[1000,1000,1000,1000,1000]
ϕ_0^{o}	veh/h	1750	3000	2000	2000
ϕ_b^{o}	veh/h	[250,250]	[500,500,500]	[500, 500, 500]	[500,500,500,500,500]
γ_b	-	[1,1]	[.25-1,.25,.25]	[1,1,1]	[1,1,1,1,1]
α_b	-	[1,1]	[.5,.5,.5]	[.15, .15,.15; .55,.55,.55]	[.2,.1,.5,.4,.15]
s_r	veh	50	50	50	50
s_b	veh	[50,50]	[100,100,100]	[0-400,0-400,0-400,0-400]	[70,100,50,40,60]
q_r^{f}	veh/h	2000	2500	2000	2000
q_b^{f}	veh/h	1000	[1200,1200,1200]	[1250,1250,1250]	[1250,1250,1250,1250,1250]
$b \in B^{\mathrm{c}}$	1/0	[0,0]	[1,1,1]	[1,1,1]	[0/1,0/1,0/1,0/1,0/1]

common, which means that the bottleneck delay can become very high if no control would be applied. This in turn implies that also the benefits of postponing the capacity drop become very high.

Figure 4.9b illustrates the change in delay given a peak period duration and a metering duration of the ramp. For instance, the benefits of ramp metering (reduction in bottleneck delay) during the first metering minutes can become 100 veh.h per minute metering for a peak period of 3 hours. The benefits per extra metering time unit (minute) decrease with an increasing metering duration.

4.7.2 Scenario B: Impact of traffic fraction to the bottleneck

Based on Scenario B it is shown that a buffers' fraction of traffic towards the bottleneck determines whether adding the buffer into the coordination is beneficial with respect to the total system performance. To this aim, the system delays are compared for the situations where a buffer is included in the coordination or left out of it, while varying the buffer's fraction to the bottleneck. If for a certain fraction, the total system delay when coordinating the buffer becomes higher than for the uncoordinated case, a break-even fraction will become apparent that indicates to what extent it is useful to include the considered buffer in the coordination. In this case, adding a third buffer to the coordination is evaluated. The ramp's metering task on the bottleneck is kept constant (when varying the extra buffer's fraction to the bottleneck) by changing the fraction of traffic to the bottleneck coming from the freeway origin.

In Figure 4.10a, b and c the system, bottleneck, ramp and buffer delays are given as a function of the fraction of vehicles towards the bottleneck of the third intersection buffer for three different peak periods. Each graph gives the delays for the situation where the third intersection buffer respectively is coordinated (solid lines) and is not coordinated (dotted lines) with the other two intersection buffers. The following general trends can be observed:

- Total ramp delays (yellow solid and dotted lines) are increasing with an increasing fraction of traffic to the bottleneck. This is due to the fact that the higher the fraction within an intersection buffer, the higher the fraction of bottleneck-oriented traffic at the ramp, the more efficient ramp storage space is used, the longer the metering duration, and the higher the total ramp delay becomes;

- The bottleneck delay (red solid and dotted lines) decreases faster than the total ramp delay (yellow solid and dotted lines) and intersection delay (purple solid and dotted lines) increase. Hence, the total system delay (blue solid and dotted lines) decrease with an increasing fraction.

With respect to adding the third buffer for extra storage space, the following can be seen in each separate graph:

- The ramp delay (yellow solid versus dotted line) increases and the bottleneck delay (red solid versus dotted line) decreases because of the extended metering duration;

- The intersection buffer delay (purple solid versus purple dotted line) increases due to the introduction of an extra buffer in which delays are caused.

When comparing the different peak period durations, it can be seen that adding an additional buffer is more beneficial for longer peak periods:

- For short peak periods (e.g. see Figure 4.10a), the caused delays for storing vehicles are in the same order as the bottleneck delay. The extra storage delay when coordinating the third buffer (difference purple solid and dotted lines + difference yellow solid and dotted lines), is for such short peak periods larger than the bottleneck delay reduction (difference red dotted and solid line);

- For longer peak periods, the benefits become much larger than the costs of coordinating the third buffer. In Figure 4.10b the system delays for the coordinated and uncoordinated situation (intersection blue solid and dotted line) indicate a break-even fraction of 42%, defining the minimum fraction of traffic to the ramp within the third intersection buffer that makes coordinating it beneficial. For even longer peak

periods (see Figure 4.10c), coordinating the third buffer is beneficial for all fractions since the coordinated system delays are always lower than the uncoordinated system delays.

To conclude, the break-even fractions for this scenario are given in Figure 4.10d as a function of the peak period duration. The curve shows that the longer the peak period becomes, the lower the required minimum fraction becomes for a buffer to be coordinated. Moreover, the figure also illustrates that for peak periods smaller than 45 minutes, it is not beneficial to include the third buffer into the coordination, whereas for peak periods longer than 65 minutes, it is always beneficial to include the third buffer.

4.7.3 Scenario C: Impact of intersection buffer storage space

By means of Scenario C is illustrated that storing a large number of vehicles can also negatively influence the network performance. This is evaluated by determining the buffer, ramp, bottleneck and system delay when applying coordination with different combinations of buffer storage space for relatively low and relatively high fractions of traffic moving from the three intersection buffers to the ramp.

Allocating large amounts of storage space at the intersection buffers with low fractions of traffic towards the bottleneck might negatively influence the total system delay. As can be seen in Figure 4.9c, the increase in intersection buffer delay (purple solid line) when adding extra storage space becomes higher than the decrease in bottleneck delay (red solid line), indicating a maximum storage space use of 150 vehicles. When dealing with a higher fraction of traffic to the bottleneck, as can be seen in Figure 4.9d, the total system delay is minimized for a storage space use of 350 vehicles.

4.7.4 Scenario D: Optimal intersection buffer configuration

In Scenario D it is demonstrated that the approach can be used to find an optimal set of buffers for coordination, and that the solution depends on the duration of the peak period and the size of the capacity drop. To this aim, a 5 buffer connection is used, in which each buffer has another storage space

and fraction of traffic moving to the bottleneck. By evaluating all coordinated buffer configurations, the optimal one can be easily found.

In Figure 4.9e and f the optimal set of ramps into the coordination is determined for peak period durations of respectively 1 hour and 3 hours, indicated by the red diamonds. By comparing the solutions, it can be seen that adopting buffers 2 and 5 in the coordination (with low bottleneck-oriented fractions according Table I-Sc.D) becomes beneficial for longer peak periods.

4.7.5 Conclusions

In this section we have presented an approach to determine the optimal buffer configuration when storing vehicles at the intersections upstream of an active ramp metering installation to extend its metering time on the freeway bottleneck. The approach gives clear insight into the factors that influence the fundamental shape and size of the intersection buffers, ramp, bottleneck and system delay. Freeway bottleneck delay can become very high, because the duration of the peak period and the size of the capacity drop have a more than linear effect on its amount. This also implies that the benefits of ramp metering can be very high at the start. However, it is also important to realize that the benefits per extra metering unit of time decrease with an increasing metering duration, hence the benefits of adding more space become less. Moreover, the break-even fraction, or the minimum fraction of vehicles towards the bottleneck at an assisting buffer that makes including the buffer into the coordination scheme beneficial, is strongly determined by the duration of the peak period and the size of the capacity drop. The higher the potential bottleneck delay (in case no control would be applied), the smaller the break-even fractions become, i.e. even buffers with a small fraction of traffic to the bottleneck might become beneficial in the coordination. To conclude, it is shown that storing too many vehicles might negatively influence the system delay in case the fractions of traffic within a buffer towards the bottleneck are relatively low.

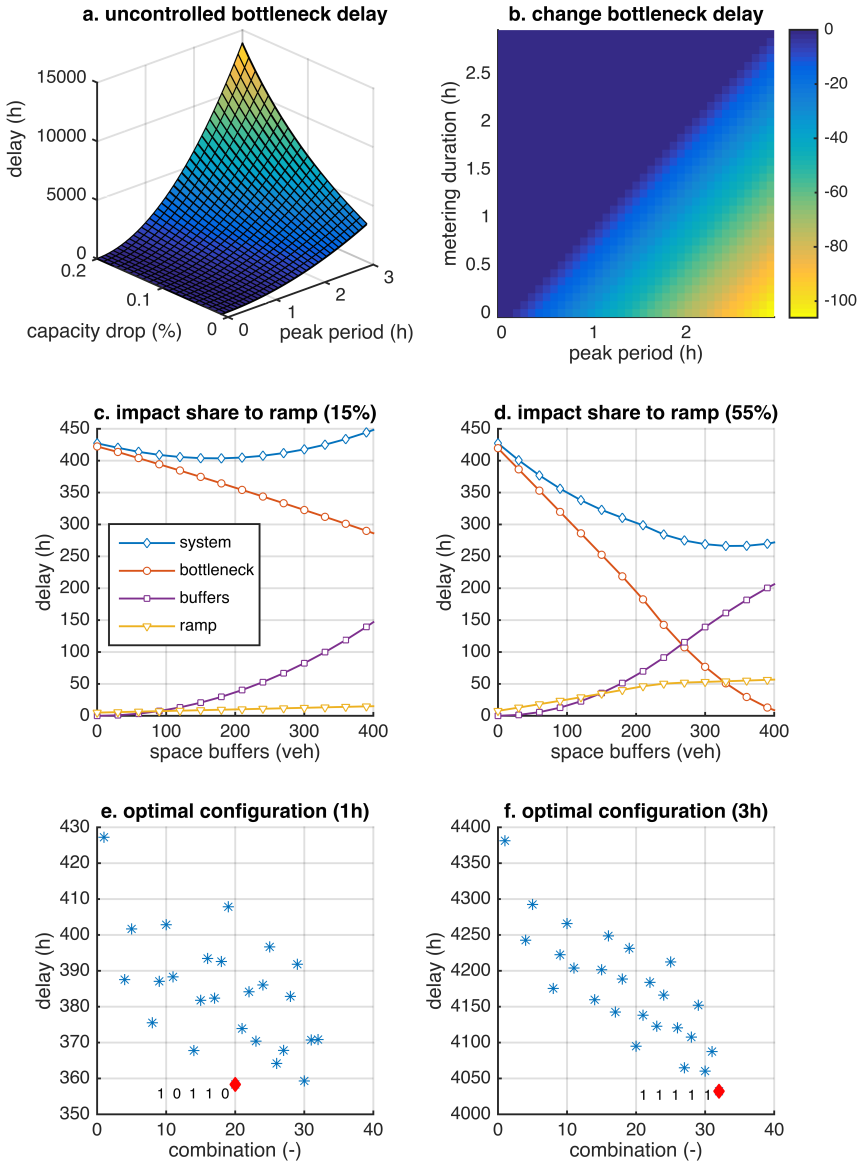

Figure 4.9: (a) Bottleneck delay as a function of the duration of the peak period and the size of the capacity drop, (b) Reduction in bottleneck delay as a function of the peak period and the metering duration, (c, d) Delays as a function of the storage space with respectively traffic fractions to the ramp of 15 and 55% for each buffer, (e,f) For peak periods of 1 hour and 3 hours, the system delay for all feasible combinations of coordinated buffers at the connection with the optimal combination indicated with a red diamond.

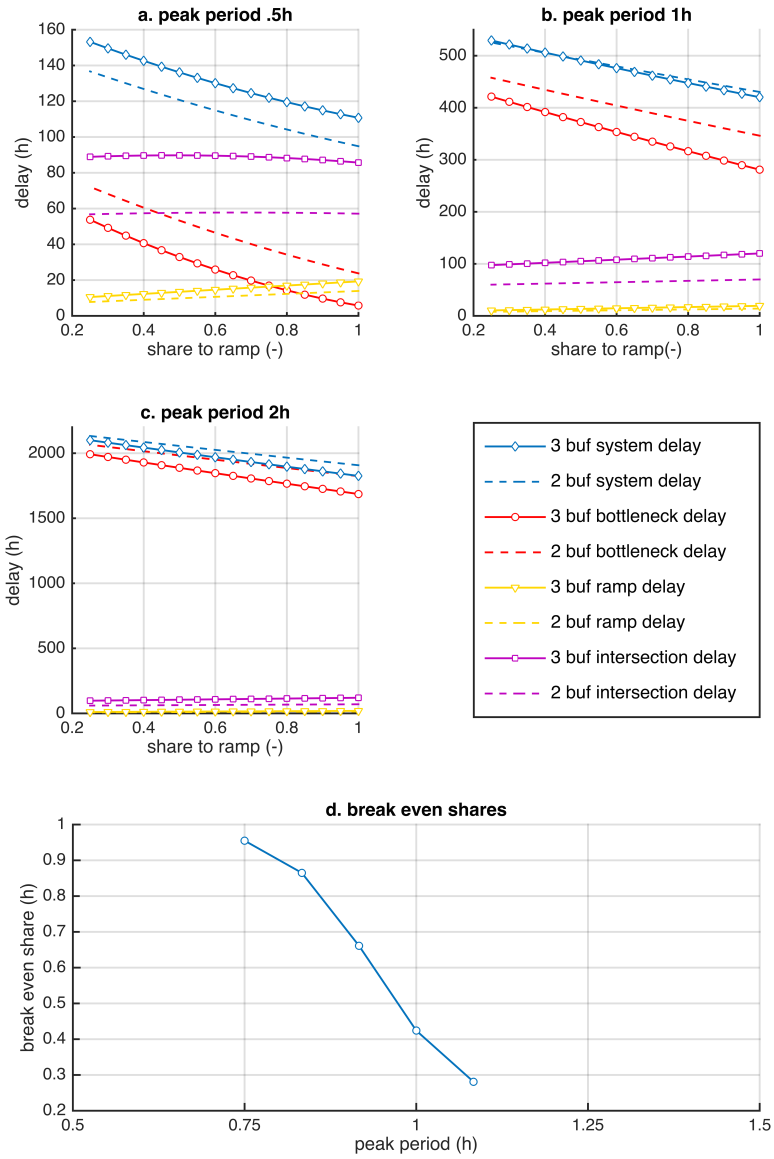

Figure 4.10: System, bottleneck, ramp and intersection delay for peak periods of (a) .5 hour, (b) 1 hour and (c) 2 hour. In (d) the break-even fractions are given for the considered situation.

4.8 Storage space filling strategies

The effectiveness with which a traffic problem is prevented is not only determined by the total amount of storage space to temporarily hold back traffic, but also by the order in which buffers are used. Filling the storage spaces can in this respect be done in a *parallel* or *serial* manner, i.e. filling all available buffers at the same time or one after the other. Both of these strategies are applied in practice by coordinated and integrated control approaches. For instance, contrary to the coordinated ramp metering scheme HERO Papamichail & Papageorgiou (2008) that realizes parallel filling of coordinated ramps, the Helper algorithm Lipp et al. (1991) employs a serial scheme. However, none of the proposed approaches discussed in Chapter 2 elaborates on their chosen filling strategy and its effectiveness.

It can be expected that network performance benefits can be achieved by filling buffers properly in a serial manner, i.e. in decreasing order with respect to the fraction of traffic towards the bottleneck. This will maximize the outflow from coordinated ramps or intersections buffers over time, because hindrance to ongoing traffic is limited.

In this section we will therefore evaluate the impact of a filling strategy by means of the previously proposed method with cumulative curves. First, a basic rule is derived on how to prioritize the use of buffers in Section 4.8.1. In Section 4.8.2, this rule serves as the foundation under the proposed strategy to cluster buffers and employ them such that the storage delays are minimized. The difference between the filling strategies is illustrated by means of a test case in Section 4.9, followed by the conclusions in Section 4.9.3.

4.8.1 Prioritization of buffers

In this section, it is shown in what order buffers should be filled during a coordination process in order to minimize delays. This evaluation is based on a widely accepted way of modeling node dynamics in macroscopic traffic flow models Tampère et al. (2011). Nodes in a traffic flow model connect one or more incoming links with one or more outgoing links. Node models are then used to compute the propagation of traffic flow between the incoming and outgoing links, while phenomena are taken into account such as queue spill back. This does enable us to model the negative effects of queue spill

back at a connection as a single node with multiple outgoing and incoming links. The outgoing links in our case would represent the ramp and ongoing directions, and incoming links would represent the feeding intersection arms.

Let us consider node n with incoming links $i \in I_n$ and outgoing links $j \in J_n$. Traffic flows from every link i to every link j are denoted by q_{ij}, which further implies that the total flow within an incoming link is given by

$$q_i = \sum_{j \in J_n} q_{ij}, \tag{4.32}$$

and that the total flow within an outgoing link is given by

$$q_j = \sum_{i \in I_n} q_{ij}. \tag{4.33}$$

The turn fractions γ_{ij}^n define for every incoming link i the partial demands d_{ij} to an outgoing link j as a fraction of the total link demand d_i

$$d_{ij}^n = \gamma_{ij}^n d_i. \tag{4.34}$$

The actual flows that move from the one to the other link can be constrained by supply and control characteristics. Phenomena like blocking back to upstream nodes can cause an outflow reduction α_j^n on the flows that do not need to pass the bottleneck (being in this case the ramp). All outgoing links $j \in J_n$ of a node n can in that respect impose an outflow reduction on the incoming links $i \in I_n$. Notice that when outflow link j^* blocks its upstream node, the supply of the congested link $R_{j^*}^n$ becomes equal to the capacity of its downstream bottleneck

$$R_{j^*}^n(k) = C_{j^*}^n. \tag{4.35}$$

For each outgoing link $j \in J_n$ of node n, an outflow reduction factor α_j^n can then be defined by

$$\alpha_j^n = \frac{R_j^n}{\sum_{i \in I^n} d_{ij}^n}, \tag{4.36}$$

indicating the ratio between the available supply R_j^n downstream and the demand $\sum_{i \in I_n} d_{ij}^n$ at node n towards j. If several links j impose a reduction on a link i, the smallest α_j^n determines all q_{ij} from link i. For each incoming link i there is thus one outgoing link j that imposes the strongest supply constraint on i. This can be written as

$$\alpha_{i^*}^n = \min_{j \in \{j | d_{ij} > 0\}} \alpha_j^n. \tag{4.37}$$

With the reduction factor for each link known, the outflow q_i of each incoming link i of node n is given by

$$q_i = \alpha_{i^*} d_i, \tag{4.38}$$

and the partial outflow q_{ij} by

$$q_{ij} = \alpha_{i^*} d_{ij}. \tag{4.39}$$

The actual flows in the outgoing links can then be written as

$$q_j = \sum_{i \in I^n} \alpha_i d_{ij}, \tag{4.40}$$

making the total outflow Q_n of the node

$$Q_n = \sum_{j \in J^n} \min(q_j, C_j). \tag{4.41}$$

Based on this model it is possible to identify the basis for an optimal heuristic strategy that maximizes the node outflow or minimizes the time spent of vehicles passing through the node. To this aim, it is assumed that the outflow reduction factors for the incoming links α_i can be controlled to prevent the negative effects of queue spill-back in the outgoing links of the node. A typical example would be a connection where the ramp is metering on a freeway bottleneck. The ramp metering rate constraints the capacity of the ramp and the outflows coming from the upstream intersection buffers can be reduced to prevent the ramp's saturation. This situation is illustrated in Figure 4.11.

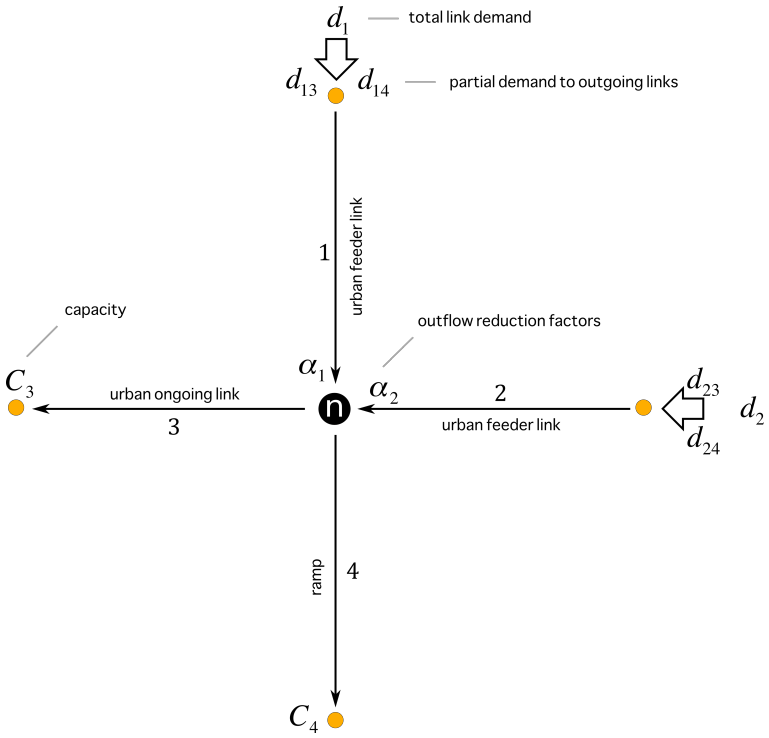

Figure 4.11: Situation used for analytical derivation optimal strategy.

The connection is modeled as a single node consisting of two incoming and two outgoing links. The incoming links $i = \{1,2\}$ represent two intersection arms where vehicles can be temporarily stored and the outgoing links $j = \{3,4\}$ represent the ongoing direction and the ramp. The capacity of destination 3 (ongoing direction) is assumed infinite $C_3(k) = \infty$, and that of destination 4 (the ramp) restricted to capacity C_4. The demand to the ramp is chosen larger than its metering rate, causing spill-back towards the incoming links of the node. The demands at the origins and the turn fraction towards the various destinations are assumed to be known and constant over time. The aim is to subsequently find the reduction factors for the outflows of link 1 and 2 that maximize the total node outflow

$$Q_n = \alpha_1 d_{13} + \alpha_2 d_{23} + C_4, \tag{4.42}$$

while satisfying the following capacity constraint to prevent spill-back

$$q_4 = \alpha_1 d_{14} + \alpha_2 d_{24} = C_4. \tag{4.43}$$

To solve the problem, (4.43) is reformulated and substituted into (4.42), resulting in

$$\alpha_1 = \frac{C_4 - \alpha_2 d_{24}}{d_{14}} \tag{4.44}$$

and

$$Q_n = [d_{23} - \frac{d_{13}}{d_{14}} d_{24}]\alpha_2 + C_4(\frac{d_{13}}{d_{14}} + 1). \tag{4.45}$$

To prevent infeasible solutions, the range of the reduction factors for the incoming links i is constrained by

$$0 \leq \alpha_i \leq \min(1, \frac{C_4}{d_{i4}}). \tag{4.46}$$

The total node outflow Q_n is then optimized by

$$\alpha_2 = \begin{cases} \min(1, \frac{C_4}{d_{24}}) & \text{if } d_{23} - \frac{d_{13}}{d_{14}} d_{24} \geq 0 \\ 0 & \text{if } d_{23} - \frac{d_{13}}{d_{14}} d_{24} < 0. \end{cases} \tag{4.47}$$

Taking a closer look to the conditions, we can formulate them as follows

$$\alpha_2 = \begin{cases} \min(1, \frac{C_4}{d_{24}}) & \text{if } \frac{d_{23}}{d_{24}} \geq \frac{d_{13}}{d_{14}} \\ 0 & \text{if } \frac{d_{23}}{d_{24}} < \frac{d_{13}}{d_{14}}. \end{cases} \tag{4.48}$$

The conditions in (4.48) can be interpreted as the ratio between the flow that is ongoing and the flow that turns to the ramp. Hence, in case there are no storage space restrictions, the connection outflow is maximized by reducing the outflow of the buffer with the highest fraction of vehicles to the ramp.

However, in practice storage space size is limited and it is not considered fair to store traffic in a single buffer while traffic from other directions is temporarily delayed. The question then rises in what order the buffers should be used to minimize the caused delays. As discussed in Papageorgiou et al. (2003), there is a direct relation between the total time spent of vehicles in a network and the realized network outflow, i.e. the network delays are minimized if the outflow is kept as long as possible as high as possible. Based on this conclusion, it can be stated that storage spaces should be used in decreasing order with respect to their fraction of traffic to the ramp (such that hindrance to ongoing vehicles remains limited as long as possible).

In the next section we will elaborate further on this concept, based on the situation where traffic is stored in the upstream intersection buffers of a ramp to prevent ramp queue spill-back to the urban network. Note that a similar approach can be followed when employing coordinated ramp metering.

4.8.2 Sequential filling strategy

The main idea of the approach proposed here, is to apply the buffers in decreasing order with respect to the fraction of vehicles moving to the ramp. It is important that enough traffic is held back to stabilize the ramp queue. To this aim, buffers might need to collaborate and be grouped into a cluster when the required outflow reduction or metering task can not be realized by an individual buffer.

To minimize the hindrance to ongoing vehicles at the connection, a cluster consists of the minimum number of buffers required to execute the metering task on the ramp. This metering task is distributed over the buffers

within a cluster to saturate them at the same time. However, simultaneous saturation is not feasible when buffers are not able to realize their given task (e.g. when the task is higher than the buffer's demand towards the ramp). The remaining task (with respect to the buffer's maximum task) is then re-distributed over the other buffers within the cluster. This in turn results in the situation where, at the end of a cluster's metering period, space remains in those buffers that were not able to meet their initial task. To fully utilize their space later on, these buffers are adopted in the next cluster.

Next sections will elaborate on how the clusters are formed including the corresponding buffer outflows. The procedure is given in Figure 4.12. Once the clusters are determined, cumulative inflow and outflow curves are defined (for the locations where delays occur) to compare the performance of the sequential and parallel filling strategy.

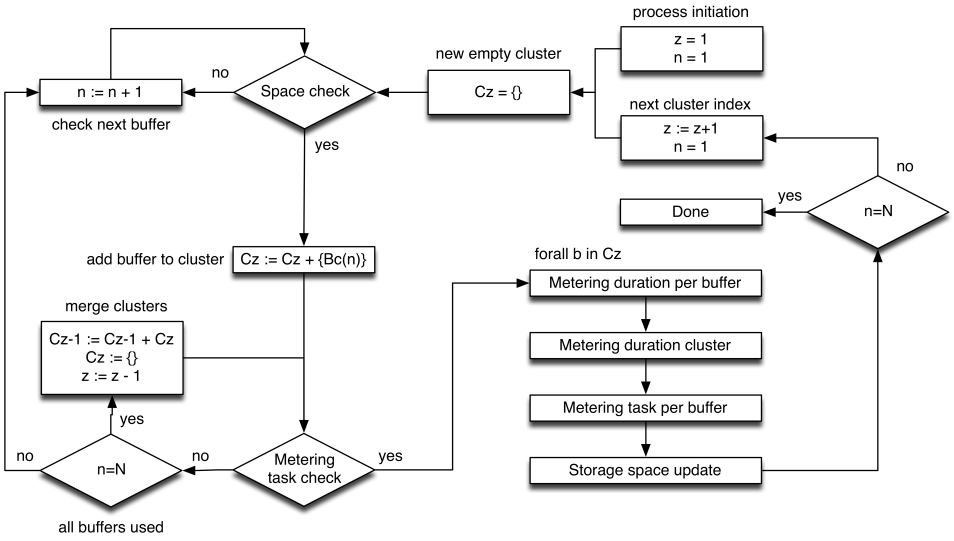

Figure 4.12: Flowchart to determine the clusters.

Step 1: Ordering buffers.

To apply the buffers such that hindrance to ongoing flows is minimized, all coordinated buffers $b \in B^{\mathrm{c}}$ are decreasingly ordered with respect to the fraction of traffic to the ramp and indexed $n = \{1,..,N\}$ with N the total number of coordinated buffers at the connection.

Step 2: Defining the clusters.

Clusters C_z become indexed $z = \{1,...,Z\}$ with Z the total number of clusters that will be formed. Clusters are initiated empty $C_z = \varnothing$ before adding the first buffer with space available from the ordered set

$$C_z := C_z + B^{\mathrm{c}}(n). \tag{4.49}$$

The metering task can be divided over the buffers within the cluster, if they are able to fulfill the required metering task

$$\sum_{b \in C_z} q_b^{\mathrm{maxtask}} > q_{\mathrm{ic}}^{\mathrm{task}}, \tag{4.50}$$

else the next buffer $n+1$ from the ordered set with space available is added to the cluster. If all buffers are already allocated to a cluster $n > N$, and an additional buffer would be needed to fulfill the metering task, the current cluster and previous cluster are merged together in order to enable maximum use of storage space

$$C_{z-1} := C_{z-1} + C_z \tag{4.51}$$

and

$$C_{z-1} := \varnothing. \tag{4.52}$$

Step 3: Calculating the buffer outflows.

If possible, the metering task is divided such, that the involved buffers within the active cluster are saturated at the same moment. To this aim, the procedure presented in Section 4.3.3 is followed. Note that the procedure is repeated for each cluster, resulting in the following notational changes:

- All intersection buffers that are active within the considered cluster are gathered in set $C_z \subset B^{\mathrm{c}}$;

- Buffers that are and are not able to realize their metering task are defined per cluster, i.e. $A_z^1 \subset C_z$ and $A_z^0 \subset C_z$, initially defined as $A_z^1 = C_z$ and $A_z^0 = \varnothing$;
- The saturation time $T_{z,b}$, the individual metering task $q_{z,b}^{\text{task}}$ and the actual metering rate $q_{z,b}$ are determined per buffer for every cluster, with $s_{z,b}$ the space in terms of number of vehicles at the start of the metering period of cluster C_z;
- The total coordination duration T_z^{c} per cluster C_z is then determined by the summation of the saturation times over all buffers that are able to realize their metering task within the cluster.

Buffer outflows $q_{z,b}$ are realized as long as the considered cluster z is active, i.e. for a period of T_z^{c}. This in turn enables us to update the available storage spaces at the end of the clusters metering period to see if buffers are available for upcoming clusters

$$s_{z+1,b} = s_{z,b} - (q_{b,z}^{\text{task}}/\alpha_b)T_z^{\text{c}}, \tag{4.53}$$

with $s_{1,b}$ the buffers' storage space at the start of the coordination period of the intersection. The moment that the ramp and all coordinated intersection buffers are filled, is determined by the saturation time of the ramp and the total metering duration of the clusters

$$t^{\text{c}} = T_r + \sum_{z \in Z} T_z^{\text{c}}. \tag{4.54}$$

From this point on, the clusters are fully identified by the following characteristics:

- Selection of buffers within each cluster;
- Buffer outflows during active clusters (to stabilize the ramp queue and use all space);
- Begin and end times of each active clusters;
- Buffers' storage space at the beginning and end of the cluster;
- Total duration that ramp saturation can be prevented.

Knowing the outflows of the buffers during each active cluster enables the definition of the cumulative inflow and outflow curves per buffer and per complete connection to determine the corresponding delays.

4.8.3 Cumulative curves for sequential filling strategy

In Figure 4.13a-c the basic shapes for the inflow and outflow curves are given for the situation where storage space runs out during the oversaturated peak period, and in Figure 4.13d-f for the situation where the off-peak period starts when there is still storage space available. These two situations are distinguished, because the relative benefits of the sequential approach are typically larger in the latter case. Since we have elaborately discussed the construction of the curves for the parallel filling strategy (in 4.4), here the focus has been put on typical aspects for building the curves for the sequential filling strategy. To clearly illustrate the differences between both strategies, two intersection buffers are adopted in the coordination where the traffic fraction to the bottleneck of buffer 1 is higher than that of buffer 2. In other words, vehicles are more effectively stored in buffer 1.

With respect to the formalization of the curve it is important to notice the following: The sequential filling strategy applies multiple clusters of buffers that realize the required outflow reduction in sequence, as can be seen in Figure 4.13a and b by the red solid line between t_r and t^{c}. The corresponding outflows during the clusters are labelled $c_{z,b}$, with z the cluster index and b the specific buffer. The slopes of the lines $c_{z,b}$ are determined by the controlled outflows $q_{z,b}$ of the buffers and determined by means of Equations (4.4) to (4.9). The slopes of lines $c_{z,b}$ differ when the *fraction of vehicles that are not moving to the bottleneck* differ per buffer. As discussed in Section 4.4, the buffer inflow over time is defined by

$$q_b^{\mathrm{in}}(t) = \begin{cases} \phi_b^{\mathrm{p}} & \text{if } t \leq t^{\mathrm{p}} \\ \phi_b^{\mathrm{o}} & \text{if } t > t^{\mathrm{p}}, \end{cases} \tag{4.55}$$

and the buffer outflow by

$$q_b^{\mathrm{out}}(t) = \begin{cases} q_b^{\mathrm{in}}(t) & \text{if } 0 < t \leq t_r \\ q_{z,b} & \text{if } t_r < t \leq t^{\mathrm{c}} \forall z \in Z \\ q_b^{\mathrm{f}} & \text{if } t^{\mathrm{c}} < t \leq t_b^{\mathrm{f}} \\ q_b^{\mathrm{in}}(t) & \text{if } t > t_b^{\mathrm{f}}. \end{cases} \tag{4.56}$$

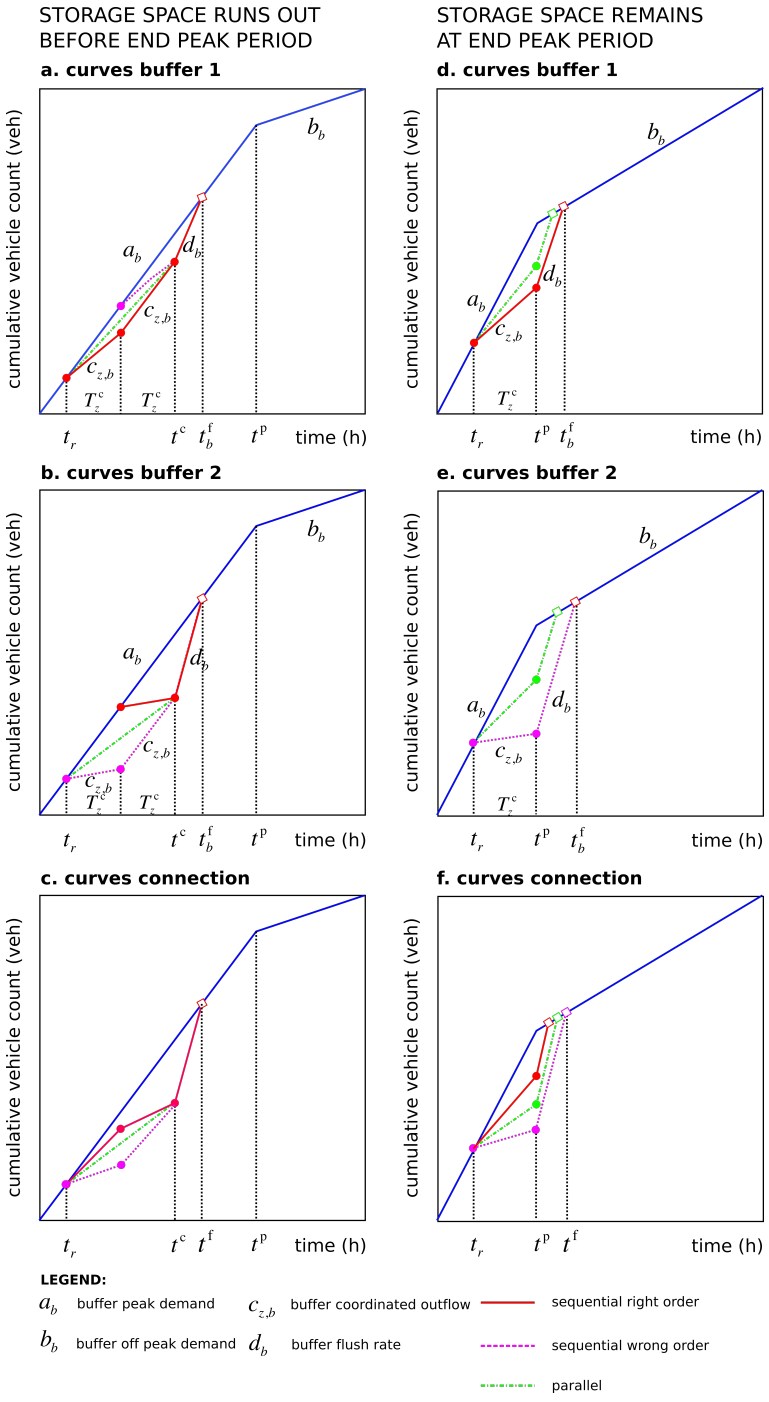

Figure 4.13: Cumulative inflow and outflow curves for two intersection buffers and the complete connection given the strategies 1. Sequential filling in the most efficient order, 2. Sequential filling in least efficient order, and 3. Parallel filling.

The curves form a multi-angle area representing the delay D_b^{tot} that is generated at the considered intersection buffer $b \in B^{\text{c}}$. The corresponding delay is determined in line with the procedure discussed in Section 4.5.

In the remainder of this section, the shape of the curves is more elaborately discussed. The red solid lines indicate the buffer outflows in case the buffers are filled in the efficient order (first buffer 1 and then 2). The magenta dashed lines indicate the buffer outflows when filling them in the least efficient order (first buffer 2 and then 1). The green dashed lines indicate the buffer outflows when filling buffers in parallel.

When using sequential filling, no delays are caused at buffers which are included in clusters that did not become active yet. Moreover, the lower the buffer's traffic fraction to the ramp, the lower its required outflow to stabilize the ramp queue (see the slopes of the different lines $c_{z,b}$). The sooner such buffers are used, the larger their impact on the buffer delay or the larger the area between the curves becomes.

Figure 4.13c and f show the aggregated buffer delays for the different strategies over time. The red, green and magenta lines show possible overall delay differences between the strategies. It can be seen that:

- Filling buffers sequentially in the right order minimizes the total buffer delay with respect to the other strategies. In Figure 4.13c it can be seen how the average outflow of the sequential strategy (red solid line) is kept as long as possible as high as possible;
- Delay gains are only achieved during the coordination period $t_r < t < t^{\text{c}}$. If the overall system delay increases due to a longer peak period t^{p}, then the relative benefits of sequential filling strategy will decrease;
- The relative benefits of the sequential filling strategy increase when the oversaturated time period t^{p} ends at the moment there is still storage space left. As can be seen in Figure 4.13f, filling the buffers in the right order also generates benefits from the moment on that vehicles are released for $t > t^{\text{p}}$, i.e. the cumulative number of vehicles that left the buffers at t^{p} is largest for sequential filling strategy.

4.9 Test case: Filling strategies

The proposed model can now be used to provide insight into the performance of the different filling strategies. In the remainder of this section the aspects are discussed that will be analyzed by means of different test case scenarios in line with the situation description given in Figure 4.2. The following questions are explored by means of the proposed approach:

- **Scenario A.** What is the impact of the peak period duration on the effectiveness of the different strategies?
- **Scenario B.** How does the traffic demand distribution over coordinated buffers impacts the effectiveness of the sequential filling strategy?

The modeled connection consists out of two coordinated intersection buffers. One has a high and the other a low traffic fraction to the ramp. Their characteristics are chosen such that they can individually realize the metering task on the bottleneck. In case of the sequential filling strategies, there are two clusters formed, each consisting of a single buffer. A detailed overview of the scenario parameter settings is given in Table 4.3. As holds for the other test cases, parameter values are chosen such that they make a meaningful scenario with interpretable results.

Table 4.3: Set-up of the test case scenarios.

Param	Unit	Scenario A	Scenario B
t^{p}	h	[.25-1]	[0.25-1]
f^{cdrop}	%	10	10
$q_{\mathrm{bn}}^{\mathrm{task}}$	veh/h	400	400
ϕ_0^{p}	veh/h	4000	4000
ϕ_b^{p}	veh/h	[600,600]	[600,600]
ϕ_0^{o}	veh/h	2000	2000
ϕ_b^{o}	veh/h	[300,300]	[300,300]
s_r	veh	[50]	[50]
s_b	veh	[50,50]	[50,50]
α_b	-	[1,1]	[.5-1,.5-1]
γ_b	-	[1,1]	[1,1]
q_b^{r}	veh/h	1500	1500
$b \in B^{\mathrm{c}}$	1/0	[1,1]	[1,1]

4.9.1 Scenario A: Impact of peak period duration

This scenario evaluates the buffer delays as a function of the peak period for the different filling strategies as can be seen in Figures 4.14a to d. Given the strategy, it can be seen how the delays are generated over time at each separate buffer. Moreover, delay at a connection level show that the parallel filling is an average strategy with respect to filling buffers in the right or wrong order. The red circles along the curves indicate the begin and end of each cluster.

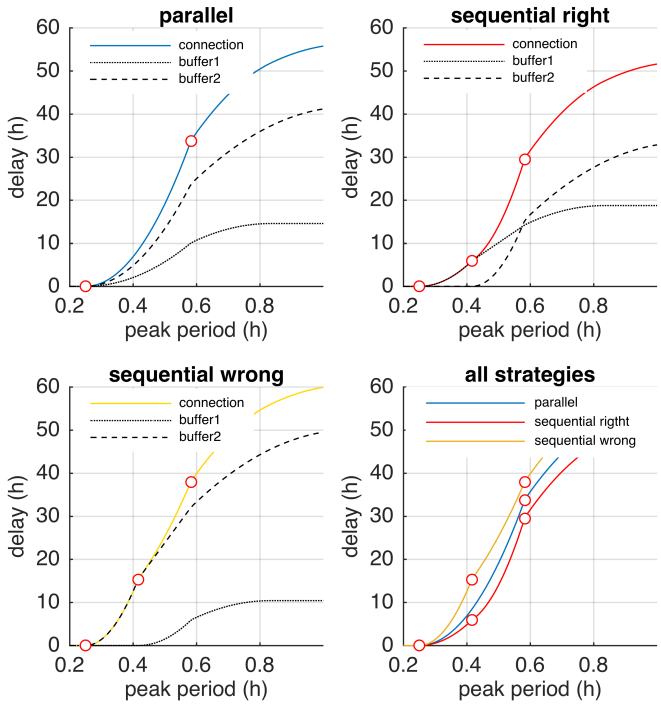

Figure 4.14: Delays due to the different filling strategies as a function of the peak period duration, with (a) Parallel filling, (b, c) Sequential filling in the right and wrong order, and (d) The performance overview of the different strategies.

Figure 4.15a and b show that applying the sequential filling strategy can be beneficial. In case the coordinated buffers do not run out of space during the peak period, 30% and 60% delay savings can be realized. These benefits

decrease as the peak period increases. The absolute and relative gains of the sequential strategy over the other strategies are indicated by the blue and red lines. Note that the blue lines define the difference between the delay curves from Figure 4.14d. With respect to the absolute delay differences:

- The delay difference increases when active clusters have a higher (average) traffic fraction to the ramp than that of the strategy it is compared with. As can be seen in Figure 4.15a and b, the maximum absolute difference is reached during the second cluster;
- When the peak period exceeds the total coordination time (i.e. $t^c = 0.58h$), the absolute difference between the strategies does not change anymore. This is due to the common point the curves share at t^c in Figure 4.13c, after which the strategies no longer differ in absolute performance.

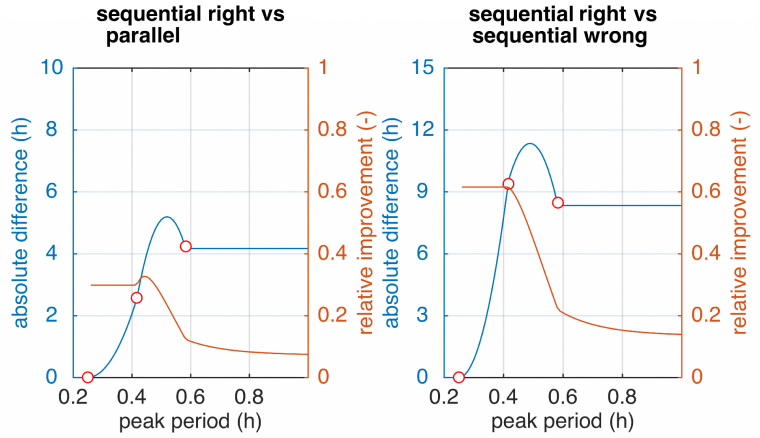

Figure 4.15: Delay difference between the Sequential filling strategy right order with (a) the Parallel filling strategy, and (b) Sequential filling strategy wrong order.

4.9.2 Scenario B: Ramp demand distribution over buffers

The benefits of filling the buffers in the right order also depends on how the fractions towards the ramp vary over the coordinated buffers. If the fractions are more or less the same, then the average hindrance to ongoing traffic

within each buffer will be the same. This implies that the hindrance over time is approximately similar to that of the parallel filling strategy. If the fractions differ significantly, it is probably more beneficial to fill buffers sequentially.

To illustrate this matter, the delays at the connection are determined as a function of the traffic fractions to the ramp of both coordinated buffers. The buffers have the same inflow and together generate a constant demand towards the ramp. The demand distribution from both buffers to the ramp is subsequently varied from 50-50% to 0-100%.

In Figure 4.16 the resulting connection delay is shown as a function of this demand distribution. As the distribution starts deviating from the 50-50% distribution, the sequential filling strategy reduces the connection delay. When there is a 80-20% distribution, the metering task can no longer be executed by the buffer with 20% fraction to the ramp. Hence, both buffers need to be grouped into a single cluster to utilize all space, which makes the sequential strategy equal to the parallel strategy.

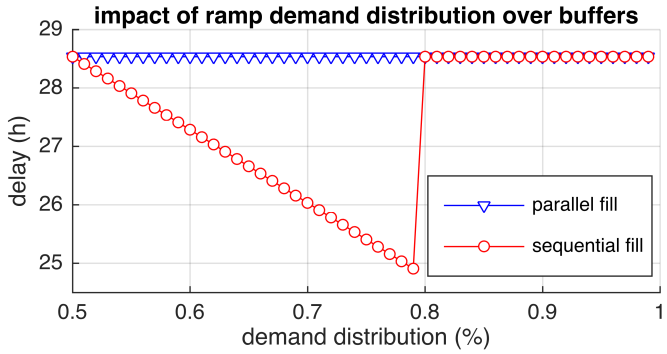

Figure 4.16: The graph shows the delay resulting from the parallel (blue line) and sequential (red line) filling strategy as a function of the demand distribution over the two buffers of traffic towards the ramp. The larger the distribution, the more beneficial the sequential strategy becomes. In this case, this holds as long as both buffers are able to individually realize the metering task.

In Figure 4.17 the performance improvement of the sequential over the parallel strategy is given as a function of the peak period duration and the ramp demand distribution over the buffers. It can be seen that applying the sequential strategy is mostly beneficial, but there are some some circumstances in which it has a negative impact. The following can be deducted from the graph:

- The start of line 1 shows that for near equal (50-50%) demand distribution and a short peak period, the sequential strategy performs worse than the parallel strategy. When all traffic is stored in a single buffer at the moment the peak period ends, it takes longer to release all stored traffic. After the break-even point (where the 'improvement' becomes positive), applying the sequential filling strategy does become beneficial;
- The steep drop in line 1 indicates the demand distribution where the metering task can no longer be executed by each individual buffer, so that the parallel strategy needs to be applied to use all space;
- Following line 2 shows that for longer peak periods, application of the sequential strategy is always preferable;
- Line 3 finally indicates that for a decreasing peak period, the improvement of the sequential strategy increases - except for short peak periods and equal demand distributions (as discussed at the first bullet).

4.9.3 Conclusions

In this section we have reflected upon a sequential and a parallel storage space filling strategy, when storing vehicles at the intersections upstream of an active ramp metering installation to extend its metering time on the freeway bottleneck. An approach is put forward to determine clusters of buffers and their outflows such that the metering task on the ramp is realized. The corresponding delays are determined by means of cumulative curves that capture the essential characteristics of the coordination approach. Insight is gained into the factors that influence the fundamental shape and size of the delays caused at the coordinated intersection buffers and the total connection.

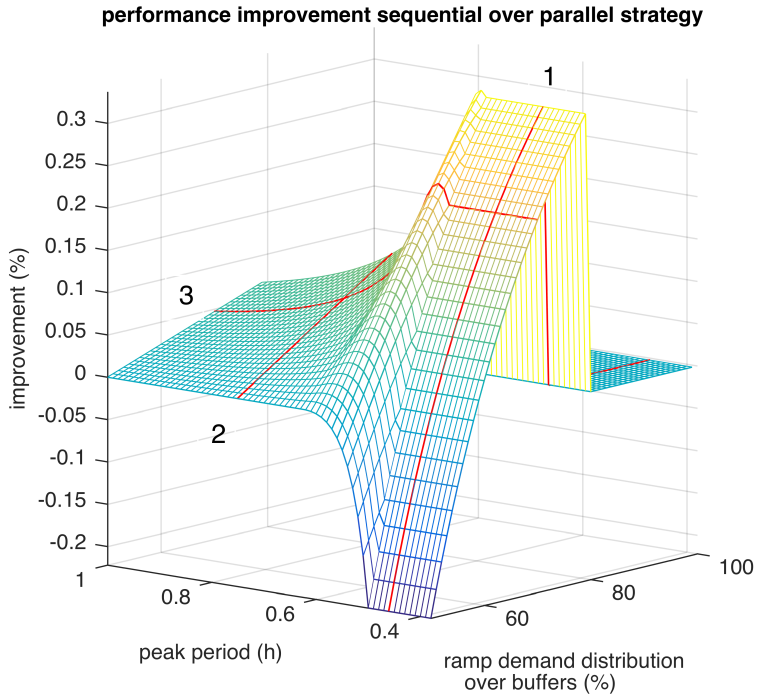

Figure 4.17: The graph shows the relative improvement that can be gained by applying the sequential filling strategy over the parallel one as a function of the peak period duration and the demand distribution over the buffers that feed the ramp. The figure shows that for shorter peak periods the improvement for this case is about 30% reduction in delay. However, if applied when the demand distribution is more or less equal (i.e. 50-50%) and the peak period short, the strategy can also result in a suboptimal performance.

By means of the sequential filling strategy large relative delay reductions can be realized over the parallel strategy. The benefits increase for shorter peak periods and a high demand distribution over the buffers that feed the ramp. The effect of the sequential strategy can also be negative in case the the demand distribution of vehicles to the ramp is more or less the same and the peak period short. If vehicles are stored in a single buffer, releasing them takes longer than would be the case for the parallel strategy. To conclude, it goes without saying that filling buffers in the wrong order has a negative effect on the delay production.

4.10 Conclusions and discussion

In this chapter we have presented an approach to determine the optimal set of buffers for coordinating and integrating traffic management measures. Its applicability has been demonstrated for coordinated ramp metering and for integrated ramp metering and intersection control. The approach is graphical, so that it gives clear insight into the factors that influence the fundamental shape and size of the involved ramp, intersection buffer, bottleneck and total system delays.

Freeway bottleneck delay can become very high, because the duration of the peak period and the size of the capacity drop have a more than linear effect on its amount. This also implies that the benefits of postponing the capacity drop by means of control can be very high at the start. However, the gains per extra metering unit of time decrease with an increasing metering duration. Hence, the benefits of coordinating more buffers and extending the metering duration become less.

The duration of the peak period and the size of the capacity drop determine wether it is beneficial to include a storage space into the coordination. The corresponding break-even fraction describes the minimum fraction of vehicles towards the bottleneck at a storage space, that makes including it into the coordination scheme beneficial. The higher the potential bottleneck delay (in case no control would be applied), the smaller the break-even fractions become. If buffers are oversaturated for a long time, buffers with a

small fraction of traffic to the bottleneck might also be beneficially included in the coordination.

It is shown that storing too many vehicles might negatively influence the system delay. This can be the case when ramps increase their metering rates to prevent spill-back to the urban roads at the end of the peak period. The bottleneck might then stay oversaturated for a longer time period (by the large amount of stored traffic flowing into the freeway) than would be the case with respect to the peak period itself. Moreover, storing too many vehicles within large intersection arms (with low fraction to the bottleneck) might cause more delay than corresponding benefits at the freeway.

We have also analyzed different storage space filling strategies that can be used to store vehicles at the upstream ramps or intersections to extend the metering duration on the freeway bottleneck. If there are large variations between buffers' fraction of traffic to the bottleneck, it is more efficient to fill buffers in a sequential way, starting with the ones that have the largest fraction. An approach has therefore been put forward to determine clusters of buffers and their outflows, such that the required metering task on the considered bottleneck is realized.

By means of the sequential filling strategy large relative delay reductions can be realized over the parallel strategy. The benefits increase for shorter peak periods and a high variation in buffers' traffic fraction to the ramp. However, the effect of the sequential strategy can become suboptimal in case fractions to the ramp are more or less equal and the peak period short. Releasing stored vehicles from a single buffer simply takes longer than from multiple buffers. To conclude, it goes without saying that filling buffers in the wrong order has a negative effect on the delay production.

Chapter 5

Coordinated ramp metering

In this chapter a coordinated ramp metering approach is proposed, based on the synchronization of the time that it takes for an on-ramp to run out of space. The duration that congestion and the associated capacity drop can be prevented at a freeway bottleneck by means of ramp metering, depends on the available ramp space for temporarily storing vehicles. For a single on-ramp, the available space and thus the metering time on the bottleneck are often limited. By means of coordination, upstream storage space can be utilized to longer prevent a breakdown. To this aim, upstream ramps need to reduce their inflow into the mainstream with respect to their own metering task. The extent to which upstream inflow reductions are effectively used to extend the metering time at the bottleneck, is determined by the order and timing at which the coordinated ramps run out of space. The proposed approach aims at saturating more upstream-located ramps always before more downstream-located ones, i.e. the ramps run out of space in downstream direction to make sure that all available storage space is used by the ramp that is metering on the bottleneck.

This chapter is based on work published in: Landman, R., A. Hegyi, S. Hoogendoorn, Coordinated ramp metering based on on-ramp saturation time synchronization, *Transportation Research Record*, Vol. 2484, pp. 50-59, 2015

5.1 Introduction

Contrary to the application of route guidance that was discussed in Chapter 3, ramp metering is a more adequate means to prevent congestion and the associated capacity drop Cassidy & Bertini (1999); Hall & Agyemang-Duah (1991) at the freeway by temporarily storing traffic at its on-ramps located along side. Such capacity drop means that the congestion after a breakdown has an outflow that is 5 to 15% lower than the free-flow capacity, depending on the severity of the congestion upstream of the bottleneck Yuan et al. (2015). Therefore, a breakdown results in suboptimal network outflow and thus a higher total time spent of vehicles in the network Papageorgiou & Kotsialos (2002).

The time congestion can be prevented at a freeway bottleneck by means of ramp metering (i.e. the metering duration) depends on the available ramp space for temporarily storing vehicles and the metering rate. In practice, local ramp storage space and thus the metering duration on the bottleneck are often limited. This is due to ramp management strategies that prevent spill-back of on-ramp queues towards the urban road network Spiliopoulou et al. (2010). At the moment a queue detector at the upstream end of the ramp becomes occupied, the ramp metering installation typically increases its ramp outflow to prevent further spill-back, causing a flow breakdown if the mainstream is near capacity.

EXAMPLE: Such limitations practically lead to rather short ramp metering durations. For instance, an on-ramp consisting of 300 meters single lane storage space (i.e. about 40 vehicles), receiving a demand towards the freeway of 700 veh/h, while realizing a metering rate of 400 veh/h, runs out of space in only 8 minutes.

By means of coordination, upstream storage space can be made available to extend the metering duration on a bottleneck at a more downstream-located ramp. To this aim, upstream ramps are controlled to reduce their inflow into the freeway. If ramps already have a local metering task, they should further reduce their outflow into the freeway. A downstream-located ramp can subsequently benefit from the reduced mainstream flow, because it will fill at a slower rate when releasing more traffic from its ramp into

the bottleneck location without causing congestion. In Figure 5.1 a more elaborate description is given on the basic principles of coordination ramp metering.

As we have seen in the background on ramp metering in Chapter 2, there have been many different algorithms proposed in literature. However, the discussion also identified that many of the earlier proposed ones as Jacobson et al. (1989); Papageorgiou et al. (1990); Lipp et al. (1991); Stephanedes (1994); Paesani et al. (1997) leave room for improvement with respect to:

- **Full utilization of ramp space.** It is important that full utilization of upstream available storage space is targeted by the coordination algorithm to maximize the metering duration on the bottleneck.
- **Keeping the bottleneck at capacity.** The bottleneck needs to be kept at capacity at all time, preferably by the ramp that is located directly upstream of the bottleneck to minimize the control delay. Shifting the metering task to upstream located ramps is likely to result in a flow breakdown. Note that in case upstream ramps have a lower fraction of traffic travelling to the bottleneck, the resulting bottleneck inflow is not sufficiently reduced to prevent a breakdown. Moreover, shifting a metering task to upstream would involve timing, because of the delayed control impact.
- **Applying feedback control.** A state feedback-based algorithm is preferred over feedforward-based types to be certain that the network state is always steered towards formulated target values that reflect desired utilization of the network and to account for unforeseen changes in traffic demand and network supply.

Despite that the coordinated ramp metering scheme HERO Papamichail & Papageorgiou (2008); Papamichail et al. (2010b) satisfies all the above mentioned aspects, the potential performance can be even further improved. HERO fills the coordinated ramps equally in order to saturate them at the same time. Notice that in case the most downstream ramp (i.e. the one in need of assistance) would run out of storage space before all upstream assisting ones do - or even when all ramps are filled at the exact same time -, there is still potential space traveling downstream that remains unused to extend the metering duration on the bottleneck. Hence, the order and timing

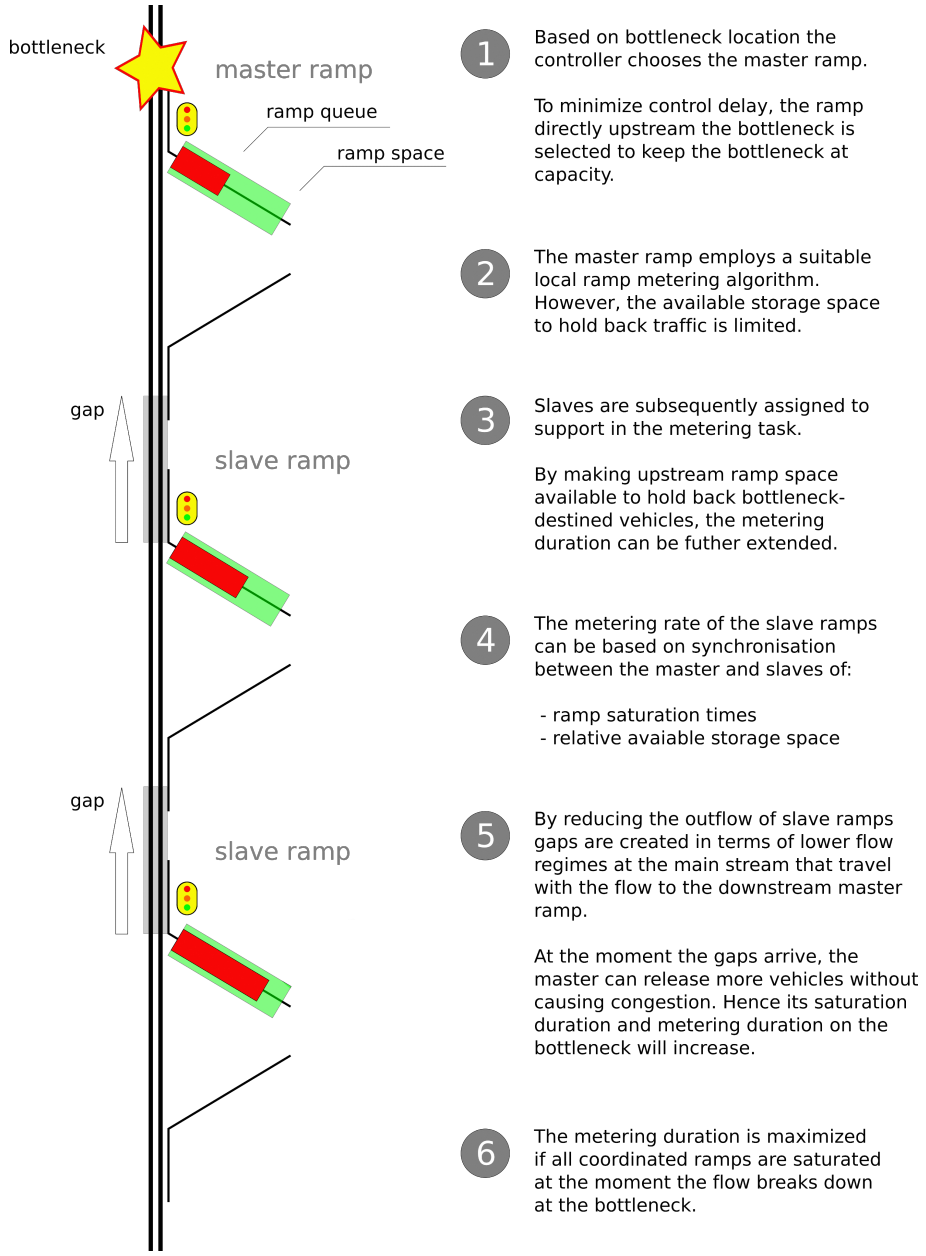

Figure 5.1: Basic working principles of coordinated ramp metering algorithms. To this aim, we consider a freeway stretch with three on and off-ramps along side. Downstream the freeway demand exceeds its supply, hence, by employing ramp metering traffic can be temporarily held back at the ramp to longer postpone a flow breakdown.

at which ramps become saturated determines whether all available space is fully and thus efficiently used.

In this chapter we therefore propose a heuristic ramp metering coordination algorithm that aims at saturating more upstream-located ramps always before more downstream-located ones, i.e. the ramps run out of space in downstream direction. This enables full utilization of upstream-realized inflow reductions of assisting ramps, to maximize the metering duration at the bottleneck and thus the time that congestion is prevented. Any local metering algorithm that adequately prevents the bottleneck location from becoming oversaturated and that aims at flow maximization can do the metering on the active bottleneck within the coordination approach.

As discussed in Chapter 4, the on-ramps to coordinate should be carefully chosen to further improve the network performance. It is shown that the selection of coordinated ramps is situation specific, i.e. depending on network, demand and control characteristics. Hence, the proposed *on-ramp selection methodology* from Chapter 4 can be used to correctly tune the performance of this (and any other) coordinated ramp metering scheme given typical conditions it encounters.

In Section 5.2, the coordination algorithm is presented, and by means of a simulation test case in Section 5.3 compared with a no-control scenario, local ramp metering and a coordination approach that aims at equalizing storage space over the coordinated ramps. To conclude, the test case results and conclusions are discussed in respectively Sections 5.4 and 5.5.

5.2 Control approach

To fully utilize the available space along the freeway corridor when a ramp meter is preventing a breakdown at a bottleneck location, upstream ramps are ordered to assist in the metering task. This coordination concept is similar to that of the well-known coordination scheme HERO and therefore the same terminology will be used for indicating the ramp that is metering on a bottleneck and for ramps that are giving assistance. In the remainder of this

contribution the ramp located directly upstream the bottleneck (in need of assistance) is called “master”, and the assisting ramps are called “slaves”.

At the slaves upstream, “gaps” are created in the mainstream flow that travel downstream with the flow towards the master. A “gap” is a region where the density (flow, and occupancy) is relatively low, created by an extra outflow reduction of the ramp with respect to its local metering task. The gaps only become fully utilized if they arrive at the master before its storage space is fully saturated. Slaves therefore need to send their final gap at least one free travel time before the master itself is expected to run out of space. To this aim, the target saturation time for the slave is synchronized with that of the master minus the free flow travel time between the slave and master.

For every slave the required outflow (slave metering rate) to meet this target saturation time can then be determined based on the available space at the considered slave ramp and the demand towards it. A slave metering rate will only be realized as long as the control constraints for minimum and maximum metering flow, and the local metering task allow for it. Furthermore, when the master is saturated while there is still storage space available on the slave ramps, the slaves need to keep assisting the master by realizing the minimum metering rate in an attempt to remove the breakdown.

If there are many on-ramps along a freeway stretch, it may occur that there are multiple active bottlenecks. In that case multiple groups of coordinated ramps might need to be formed consisting of a master and assigned slaves to best deal with the situation at hand. However, in this contribution it is assumed that the coordination algorithm will only form one coordination group. Moreover, the most downstream ramp in need of assistance is given the master status, as the corresponding bottleneck probably has the largest impact on the network outflow.

5.2.1 Coordination algorithm: the basics

To develop and test our proposed algorithm, we make use of a traffic flow simulation model with which the real-world process is replicated. In the model, time is discretized in time steps $k = 1, 2, ..., K-1$ indicating time instant $t = kT$, with K the total number of simulation time steps and T the time step size. A distinction is made between this time step size and the time

step size T_c after which the controllers are activated. To properly embed the controller within the simulation process, it is assumed that T is an integer divisor of T_c:

$$T_c = MT, \tag{5.1}$$

with M an integer. Hence, when the coordination algorithm is activated at control step k_c, the corresponding simulation time index is $k = Mk_c$. The ramp metering installations that can be coordinated along the considered freeway are indexed $r = 1, ..., r_{\max}$, with the most downstream installation $r = 1$ and the most upstream installation $r = r_{\max}$. Storage space at a ramp r can be expressed in absolute terms and in relative terms. The absolute storage space $s_r(k_c)$ is defined by

$$s_r(k_c) = s_r^{\max} - w_r(k_c), \tag{5.2}$$

with $s_r^{\max}(k_c)$ maximum storage space of ramp r and $w_r(k_c)$ the ramp queue in terms of vehicles. Relative storage space is then determined by

$$\theta_r(k_c) = \frac{s_r(k_c)}{s_r^{\max}}. \tag{5.3}$$

In the remainder of this section we elaborate on how the master and slave metering rates are determined when the coordination is active by the following subjects:

1. Activation of the coordination: assigning master and slaves
2. Metering rate of the (candidate-) master
3. Saturation time of the master
4. Target saturation time of the slaves
5. Slave metering rates
6. Realized ramp metering rates

Activation of the coordination: assigning master and slaves

The coordination algorithm receives every control interval the status of the ramp metering installations under control. When a ramp metering installation is metering on a bottleneck, the coordination algorithm gives the considered ramp a candidate-master status. This means that the ramp is considered for the master status where it receives assistance from all upstream located ramps in the coordination group. The indices of the candidate-masters are gathered in the set $A(k_c)$. Coordination is activated when at least one of the candidate-master ramps' relative storage space $\theta_r(k_c)$ has dropped below a certain threshold $\theta^{\text{threshold}}$. The most downstream ramp that meets this requirement receives the master status

$$m(k_c) = \min(r \mid \theta_r(k_c) < \theta^{\text{threshold}},\ \ r \in A), \tag{5.4}$$

and all upstream ramps become slave

$$S(k_c) = \{m+1, ..., r_{\max}\}. \tag{5.5}$$

Coordination will stop in control interval k_c when there are no more ramps meeting the requirement for becoming master in (5.4).

Metering rate of the (candidate-) master

Candidate-masters and masters are preventing a flow breakdown at a bottleneck location. The corresponding local metering rate $q_r^{\text{loc}}(k_c)$ can be determined by any suitable local ramp metering algorithm, like ALINEA Papageorgiou et al. (1991).

Saturation time of the master

The main concept behind determination of the slave metering rates (veh/h) is that the saturation time of slave ramps (h) is synchronized with the saturation time of the master ramp (h) minus the travel time between the considered slave and master (h). The saturation time $T_r(k_c)$ of the master ramp $r = m$ is determined by

$$T_r(k_c) = \frac{s_r(k_c)}{\Delta s_r(k_c)}, \tag{5.6}$$

with $s_r(k_c)$ the storage space and $\Delta s_r(k_c)$ the change in storage space in terms of vehicles. The change in storage space is simply determined by the difference between the current storage space and that of the previous control interval

$$\Delta s_r(k_c) = s_r(k_c - 1) - s_r(k_c). \tag{5.7}$$

If $\Delta s_r(k_c)$ is constant and larger than zero (assuming a fixed metering rate and fixed demand into the ramp), then $T_r(k_c)$ describes the time that the on-ramp runs out of space. A positive saturation time (decreasing storage space) means that slaves need to assist and decrease their inflow into the mainstream. A negative saturation time (increasing storage space), results in the incentive for the slaves to also release more traffic. It is further assumed that a zero change in storage space results in an infinite saturation time (division by 0 in (5.6)). This is intuitively right, as the master ramp would then never become saturated.

Target saturation time of the slaves

The target saturation time $T_i(k_c)$ for the slaves $i \in S$ can now be determined. If the saturation time of the master $T_m(k_c)$ is positive, the target saturation time of the slaves is determined by the saturation time of the master m minus the travel time $\tau_{i \to m}$ between the master and considered slave i. Note that the travel time correction on the target saturation time forces upstream slaves to run out of space before the master does. When the master ramp runs empty it has a negative saturation time. To make the slave also release its stored vehicles, the target saturation time of the slave is set equal to this value which leads to an increased slave metering rate. This results in the following equation:

$$T_i(k_c) = \begin{cases} \max(c, T_m(k_c) - \tau_{i \to m}), & T_m(k_c) \geq 0 \\ T_m(k_c), & T_m(k_c) < 0. \end{cases} \tag{5.8}$$

Note: In (5.8) the target metering time of the slave can become negative when a positive metering time of the master is shorter than the travel time between the considered ramps. In other words, the master needs assistance because it is almost saturated, but the gaps will not arrive in time. To make slaves nevertheless meter strongly, the target saturation time set to a small constant value c (e.g. 0.01).

Slave metering rates

As long as the master has space available, the desired ramp flow for each slave $i \in S$ that realizes the target saturation time, can be determined by (5.9) based the demand into the slave ramp $d_i(k_c)$ and its available space $s_i(k_c)$:

$$q_i^{\text{crd}}(k_c) = \begin{cases} d_i(k_c) - \frac{s_i(k_c)}{T_i(k_c)}, & \text{if } s_m(k_c) \geq 0, s_m(k_c-1) > 0 \\ q_i^{\min}, & \text{if } s_m(k_c) = 0, s_m(k_c-1) = 0. \end{cases} \tag{5.9}$$

There is however an exception, a saturated master over multiple control intervals, needs to keep receiving maximum assistance to prevent or resolve a breakdown when there is still space available at the slaves. To this aim, the slave metering rate is set to its minimum metering rate $q_i^{\min}$. This exception needs to be made, because if there is no more storage space on the master in control intervals $k_c - 1$ and k_c, the change in storage space $\Delta s_r(k_c)$ would become 0 (given (5.7)). Then the target saturation time for the slave would become infinite (given (5.6) and (5.8)), resulting in the undesired incentive to equalize the slave metering rate $q_i^{\text{crd}}(k_c)$ to the demand into the ramp (according(5.9)).

Realized ramp metering rates

To guarantee that a coordination request will not lead to a local problem, we chose for every ramp meter the minimum of the local and coordinated metering rate

$$q_r(k_c) = \min(q_r^{\text{loc}}(k_c), q_r^{\text{crd}}(k_c)). \tag{5.10}$$

To conclude, the actually realized metering rate $\hat{q}_r(k_c)$ is constrained by maximum and minimum metering rates of the considered installation, respectively $q_r^{\max}$ and $q_r^{\min}$:

$$\hat{q}_r(k_c) = \min(q_r^{\max}, \max(q_r^{\min}, q_r(k_c))). \tag{5.11}$$

5.2.2 Coordination algorithm: the extension

In the above elaborated algorithm the master makes an underestimation of the available space, since gaps that are already traveling towards the master are not accounted for at the determination of the saturation time $T_m(k_c)$. As

a result, the metering time is underestimated and the slaves would have to meter harder than necessary. The determination of the saturation time of the master is therefore reformulated in such a way that we do account for the traveling gaps:

$$T_m(k_c) = \frac{s_m(k_c) + s_m^{\text{gaps}}(k_c)}{\Delta s_m(k_c)}, \tag{5.12}$$

with $s_m^{\text{gaps}}(k_c)$ an approximation of the created gaps in previous control intervals. The size of a gap that is created in a control interval is determined by

$$s_i^{\text{gap}}(k_c) = (\min(d_i^{\text{rm}}(k_c-1), q_i^{\text{loc}}(k_c-1)) - q_i^{\text{crd}}(k_c-1)) * T_c, \tag{5.13}$$

with T_c the size of the control interval, $d_i^{\text{rm}}(k_c-1)$ the traffic demand out of the on-ramp, $q_i^{\text{loc}}(k_c-1)$ the ramp flow due to local ramp metering and $q_i^{\text{crd}}(k_c-1)$ the ramp flow due to coordination at the previous control interval k_c-1. The total amount of gaps that has been sent by the slaves $i \in S$ to the master in the previous control interval k_c-1 then becomes

$$s_m^{\text{gaps}}(k_c) = \sum_{i \in S} s_i^{\text{gap}}(k_c). \tag{5.14}$$

5.2.3 Discussion on measurement errors

To maximize the metering time on a bottleneck it is important to saturate coordinated ramps in the right order and at the right moments. The different variables used to determine the slave metering rates influence the realized timing such as: the predicted saturation time of the master ramp, the demand into the slave ramps, and the assumed free travel time between the ramps. In the remainder of this section the effects on the ramp saturation order and timing are explained, in case of prediction errors in the above mentioned variables. These errors may be due to measurement errors, calibration errors, and the interdependencies between the predictions and the future control actions.

Every control interval an estimation is made of the instantaneous saturation time of the master ramp, given the available space at the master ramp, an estimate of the traveling gaps within the mainstream, and the measured

change in storage space at the master. It is, however, impossible to instantaneously determine the true saturation time, because the saturation time of the master is continually influenced by the realized slave metering rates (that are based upon the saturation time of the master). Nevertheless, the desired control behavior is realized by recalculating the saturation time of the master every control interval based on real-time measurements.

An overestimation of the saturation time due to an overestimation of the storage space at the master or an underestimation of change in storage space, might lead to suboptimal performance. The metering task for the slaves will then become smaller than required, meaning that there will remain storage space at the slave ramps at the moment the master is truly saturated. In other words, a flow breakdown occurs at the master ramp, resulting in reduced system outflow while there is still space left at the slaves.

An underestimation of the saturation time also has a negative impact on the network performance. The metering rate of the slaves will become smaller than required, meaning that slaves are saturated well before the master is. In case an upstream breakdown happens when slaves increase their metering rate after being saturated, the master might benefit from the resulting capacity drop. If the capacity upstream does not drop, the master might be able to keep the mainstream at capacity and use its own remaining space to extend the metering time on the bottleneck. On the other hand, if the resulting mainstream flow is close to capacity, a breakdown might occur at the master, leaving storage space at the master ramp unused to extend the metering time on the bottleneck.

Similar reasoning holds for over and underestimating the demand into the slave ramps. The slave metering rate is based on the demand into the slave ramp, its available space and the target metering time. An overestimation of the demand makes the slave saturate too late, and an underestimation makes the slave saturate too early.

The travel time between a slave and master also has a direct impact on the target saturation time of a slave. If the travel time is underestimated, not all traveling gaps will reach the master before it breaks down, leaving potential assistance in terms of mainstream gaps unused. An overestimate of the travel time gives slaves the incentive to saturate too early.

By means of a simulation test case that is presented in the next section, we will illustrate the impact of the measurement errors on the system performance.

5.3 Test case

The focus of the simulation experiments is to illustrate the relevance of properly timing the saturation of coordinated ramps. To this aim, a comparison is made with an algorithm that also explicitly takes storage space into account, but without the timing aspect of making slave ramps saturate earlier. This algorithm can be considered a simplified version of HERO and will be introduced in the remainder. Furthermore, to illustrate the benefits of applying coordination, comparisons are made to a no-control and a local ramp metering scenario. To conclude, the consequences are analyzed of making prediction errors on the saturation time and demand into the slave ramps.

The applied traffic network and its characteristics is given in Figure 5.2. The network consists of a freeway stretch with two on-ramps along side that are equipped with ramp metering installations. In this test case all traffic that enters the network travels towards the freeway destination in the east, meaning that all traffic coming from the ramps moves along the bottleneck location. The aim is to illustrate the benefits of making slave ramps run out of space before the master does, so that the master is able to use all upstream storage space to maximize the metering time on the bottleneck. Note, that coordination can have a negative effect on the network performance when slaves are holding back traffic that does not need to pass the master, i.e. vehicles from the coordinated ramps are leaving the mainstream at an off-ramp upstream of the master.

The demands entering the freeway are gradually increased such that a bottleneck is activated at the most downstream ramp merge area. The corresponding ramp metering installation will start metering to prevent a breakdown on the mainstream flow and the coordination algorithms will coordinate the ramps along the freeway in an attempt to maximize the metering time on the bottleneck. Note that the system performance is fully determined

by the network outflow downstream the most downstream merge. Hence, the longer the capacity drop is prevented, the better.

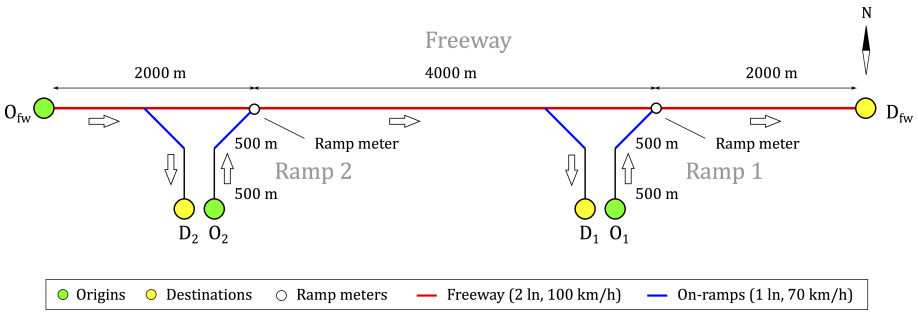

Figure 5.2: Network layout for the test case and its corresponding characteristics.

Traffic is generated at origin O_{fw} on the freeway and at the origins O_1 and O_2 at the ramps. All traffic moves towards destination D_{fw} at the freeway. The traffic demands over time are given in Table 5.1.

Table 5.1: Traffic demand loaded at origins in terms of veh/h.

Time (hh:mm)	7:00	7:15	7:30	7:45	8:00	8:15	8:30
$d_{O_{\text{fw}} \to D_{\text{fw}}}$	1750	3500	3500	3500	3500	1750	0
$d_{O_{\text{r1}} \to D_{\text{fw}}}$	100	600	600	600	600	300	0
$d_{O_{\text{r2}} \to D_{\text{fw}}}$	500	500	500	500	500	250	0

5.3.1 Applied traffic flow model

The macroscopic multi-class cell-based traffic flow model Fastlane van Lint et al. (2008) has been used for the process simulation. Fastlane is able to properly reproduce the phenomena that influence the network performance, i.e. the build up and dissolving of congestion, blocking back effects of queues and the capacity drop. Fastlane applies a Godunov-scheme to model the flows between cells. To properly replicate the capacity drop phenomenon in Fastlane, the supply of the cells located directly downstream of cells that are congested is decreased by a certain factor (e.g. 15%).

5.3.2 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 all vehicles in the network (incl. queues on the on-ramps) over a period $k = \{0, 1, ..., K-1\}$ with K the total number of simulation time steps is determined by

$$J_{\text{TTS}} = T \sum_{k=1}^{K} \sum_{m \in M} \sum_{c \in C_m} \rho_{m,c}(k) \lambda_{m,c}, \tag{5.15}$$

with $\rho_{m,c}(k)$ the vehicle densities (veh/km) over the cells $c \in C$ of all links in the network $m \in M$ and $\lambda_{m,c}$ the corresponding cell lengths (in km). The total delay (TD) for each scenario is determined by subtracting the free TTS from the TTS of each scenario. Moreover, the start and end time of congestion t_0^{cong} and t_1^{cong}, its duration $T^{\text{cong}} = t_1^{\text{cong}} - t_0^{\text{cong}}$, and maximum queue length $W_{\text{cong}}^{\max}$ are compared.

5.3.3 Benchmark algorithms

In this section the algorithms are discussed with which the proposed approach will be compared in the simulation test case.

Local ramp metering

A demand-capacity algorithm is implemented for local ramp metering. The ramp metering rate $q_r^{\text{loc}}(k_c)$ is based on the mainstream flow upstream of the ramp $q_r^{\text{up}}(k_c)$ and a capacity estimate of the road downstream the merge C_r. However, when the speed downstream the merge $v_r^{\text{dn}}(k_c)$ drops below a defined threshold v_r^{cong} (e.g. 50 km/h) the minimum metering rate is applied. The maximum metering rate is applied when the ramp queue $w_r(k_c)$ exceeds a maximum defined length $w_r^{\max}$ in terms of vehicles. The algorithm then becomes

$$q_r^{\text{loc}}(k_c) = \begin{cases} C_r - q_r^{\text{up}}(k_c), & \text{if } w_r(k_c) < w_r^{\max}, v_r^{\text{dn}}(k_c) \geq v_r^{\text{cong}} \\ q_r^{\min}, & \text{if } w_r(k_c) < w_r^{\max}, v_r^{\text{dn}}(k_c) < v_r^{\text{cong}} \\ q_r^{\max}, & \text{if } w_r(k_c) \geq w_r^{\max}. \end{cases} \tag{5.16}$$

Coordination based on synchronization of ramp storage space

This coordination strategy makes use of a feedback control approach that aims at equalizing the the available relative storage space of the slaves with that of the master. Hence, the aim of the algorithm is to equally fill the ramps during over-saturated conditions and thus that the storage space at the ramps runs out at approximately the same time. All assisting slaves receive a target value (setpoint) in terms of relative storage space, that is equal to the current relative storage space at the master

$$\theta^* = \theta_m(k), \tag{5.17}$$

with θ^* the target value for the slaves and $\theta_m(k)$ the relative storage space at the master. Subsequently the slave metering rate for every slave $i \in S$ is determined by

$$q_i^{\text{crd}}(k) = q_i^{\text{crd}}(k-1) - \alpha_1 \cdot e_i(k) - \alpha_2 \cdot \Delta e_i(k), \tag{5.18}$$

with $q_i^{\text{crd}}(k-1)$ the slave metering rate from the previous control interval, and the feedback gains for the proportional and integral terms, respectively $e_i(k)$ and $\Delta e_i(k)$. The proportional term $e_i(k)$ determines the difference between the relative storage space at the slave $\theta_i(k)$ and the setpoint θ^*

$$e_i(k) = \theta_i(k) - \theta^*, \tag{5.19}$$

and the integral term $\Delta e_i(k)$ accounts for the change in of the error over time

$$\Delta e_i(k) = e_i(k) - e_i(k-1). \tag{5.20}$$

5.3.4 Set-up of the coordination algorithms

In both approaches the master ramp will meter on the bottleneck location by means of the above introduced local metering algorithm. Coordination is activated when the relative storage space of a ramp becomes less than 95%, i.e. $\theta^{\text{threshold}} = 0.95$. The coordination approach based on synchronization of saturation time requires no further calibration. The coordination approach based on synchronizing storage space only requires the applied feedback gains α_1 and α_2 to be calibrated. Proper values are derived based structural evaluation of the system performance for different combinations of feedback

gains α_1 and α_2. The best system performance and situation where the slave storage space closely follows the storage space at the master is achieved for gains: $\alpha_1 = 200$ and $\alpha_2 = 1250$.

5.4 Results

In this section the results of the various test case scenarios are presented. We will subsequently take a closer look at the freeway conditions in terms of speed and flow, the way the control approaches utilize the ramp storage space and of course the overall network performance in terms of total time spent.

5.4.1 Space-time diagrams

In Figure 5.3a and b the space-time diagrams for speed and flow are given when no ramp metering is applied. Over 6 km of congestion builds up from 7:18 h onwards, lasting to 8:24 h (duration 66 min). This means that almost the entire simulation the capacity of the freeway was reduced by 15%.

Figure 5.3c and d show the situation when only local ramp metering is applied. From the speeds it can be seen that at 7:36 h a queue of 5 km starts building up, still spilling back over the upstream connection. The flow diagram indicates, that from then on, the capacity drop is active at the bottleneck location (km6), reducing the network outflow until 8:27 h (duration 51 min).

When applying the coordination based on equally filling the ramps, we can see in Figure 5.3e and f that the flow breaks down first at the most downstream ramp at 7:51 h followed by a breakdown at 7:54 h at the upstream ramp. The congestion lasts until 8:22 h (duration 31 min) and spills back 2.9 km.

As can be seen in Figure 5.3g and h, the order at which the ramps become saturated for the proposed approach is different, i.e. the flow at the upstream slave breaks down first. Congestion sets in at 7:54 h and lasts till 8:22 h (duration 28 min). Moreover, the spill back remains limited to 2.3 km. The time difference between the breakdown at the master and that at the slave

is about three minutes, which closely approximates the travel time between them. Hence, the final gaps created by the slave in the proposed approach all reached the master before it ran out of space.

5.4.2 Ramp usage

In Figure 5.4a it is visible that only the storage space at the most downstream ramp (ramp 1, red line) is used to meter on the bottleneck. The more upstream located ramp meter (ramp 2, blue) becomes shortly active due to the spill back of congestion (see Figure 5.3a). The oscillations after 7:30 h are caused by the ramp management strategy that is activated when the ramp is saturated.

From Figure 5.4b the feedback character of the coordination approach based on storage space synchronization is clearly visible. Note that the slave ramp (ramp 2, blue line) constantly follows the target storage space at the master ramp (ramp 1, red line). Moreover, the master becomes saturated before the slave does.

The functioning of the proposed coordination approach is elaborated based on Figure 5.4c and d. At point *A* the master starts metering on the bottleneck, resulting in decreasing storage space. The target saturation time then becomes positive, giving the slave the incentive to also reduce its inflow by offering storage space. At point *B* the gaps arrive and the ramp storage space is stabilized. Between points *B* and *C* the saturation times therefore become either very large or infinite (not plotted), allowing the slaves to also release traffic. When the higher flows arrive downstream at the master (point *C*) it needs to start metering strongly again, further reducing its storage space. At point *D* an example can be seen of the situation where the gaps are large enough for the master to release more traffic than the demand into the ramp, resulting in a negative metering time that makes the slave also release traffic.

5.4.3 Network performance

To conclude, the resulting overall network performance is given in Table 5.2. Both the coordinated ramp metering approaches realize a large improvement in the network performance with respect to the no-control case and the sit-

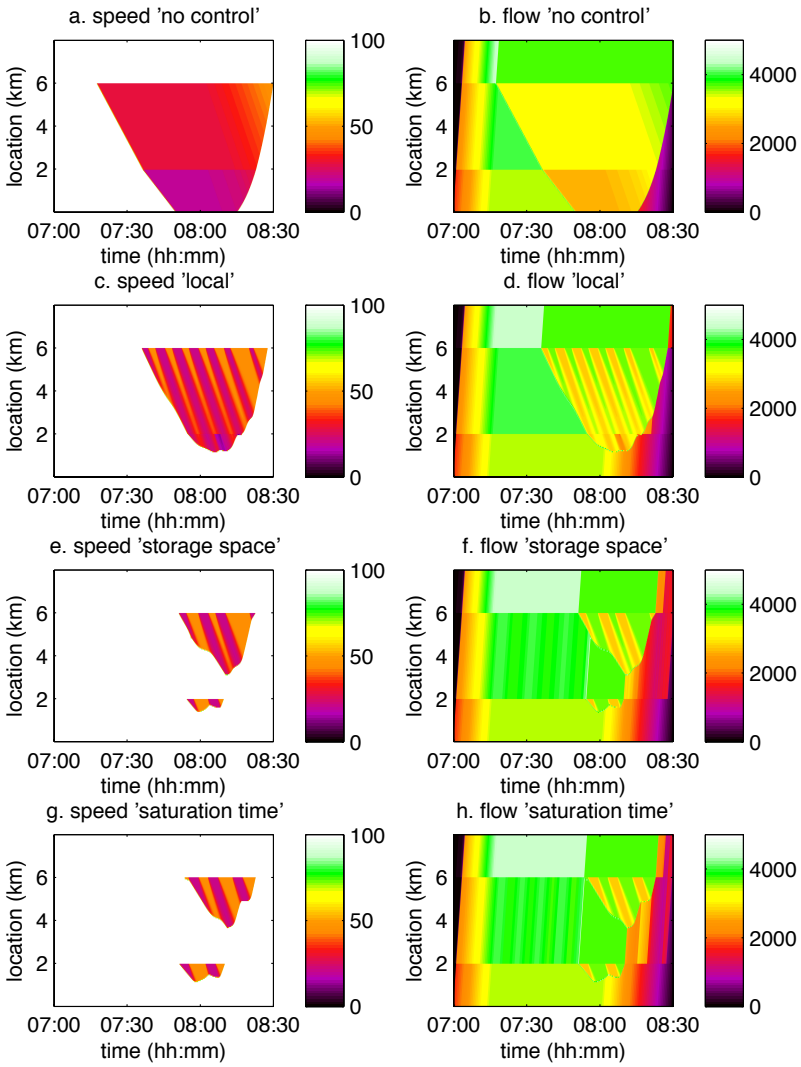

Figure 5.3: Space-time diagrams for speed and flow of no-control scenario (a,b), Local metering (c,d), Coordination based on saturation time synchronization (e,f), Coordination based on storage space synchronization (g,h).

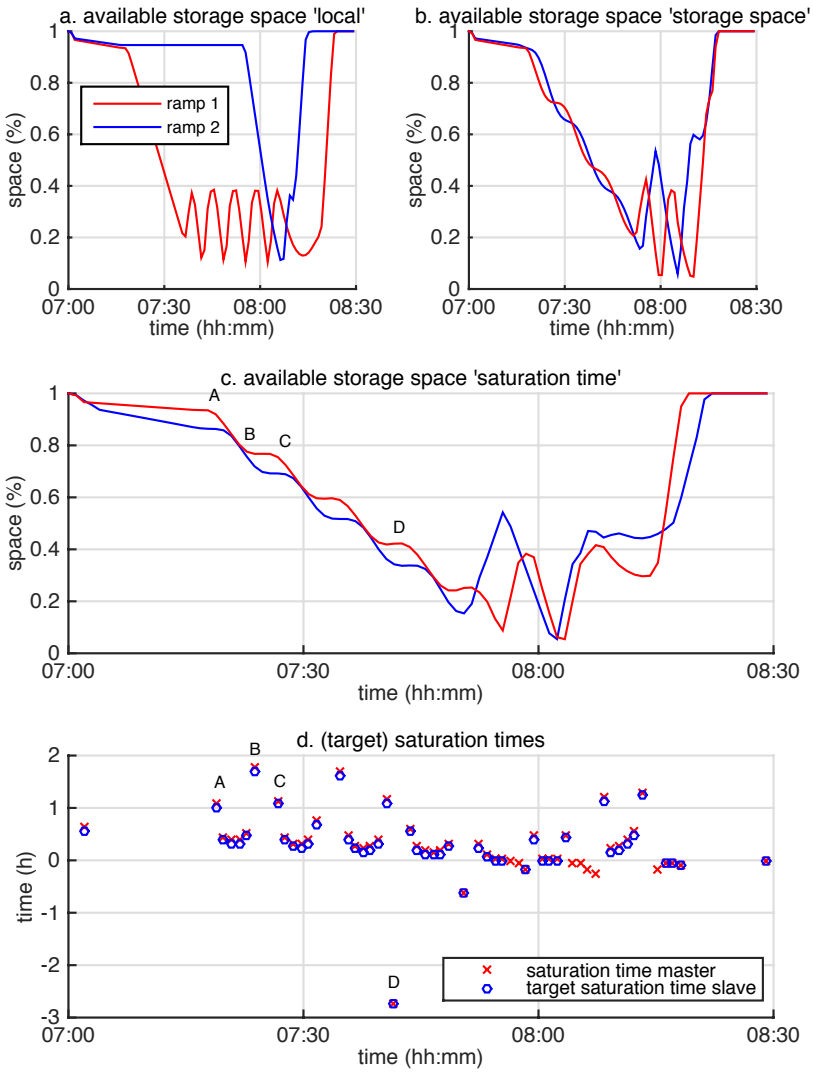

Figure 5.4: Utilization ramp storage space for (a) Local metering, (b) Coordination based on storage space synchronization, (c) Coordination based on saturation time synchronization, and (d) The corresponding (target) saturation times.

uation where only local ramp metering is applied. Coordination based on saturation time synchronization results in a slightly better network performance than the synchronization of storage spaces.

Table 5.2: Overview network performance indicators.

Approach	TTS (h)	TD (h)	t_0^{cong} (hh:mm)	t_1^{cong} (hh:mm)	T_{cong} (min)	$W_{\text{cong}}^{\max}$ (km)
no-control	830	418	7:18	8:24	66	>6
local control	691	279	7:36	8:27	51	5
coord. storage space	574	162	7:51	8:22	31	2.9
coord. saturation time	562	150	7:54	8:22	28	2.3

5.4.4 Measurement errors

In Figure 5.5a and b the effects on the network performance are shown when the saturation time of the master and the demand into the slave ramp are structurally over- and underestimated. In both graphs the network performance corresponding to error values to the right of the zero-error results in the master ramp being saturated too early, and to the left of the zero-error results in slave ramps being saturated too early. The graphs show an unique optimum, which indicates that taking the saturation timing into account is essential. Structural measurement errors in the variables that determine both the saturation time of the master and the metering rate of the slaves, will thus result in suboptimal network performance. Hence, accurate monitoring is required of the available space at the master and slave ramps, and of the demand into the slave ramps.

5.5 Conclusions

In this chapter a coordinated ramp metering approach is presented based on saturation time synchronization. The test case results show the importance of taking into account the moment of saturation of coordinated on-ramps. They also illustrate the potential of the algorithm to fully use all ramp space along a freeway when metering on a bottleneck. The proposed approach improves the network performance with respect to a coordination approach that aims at

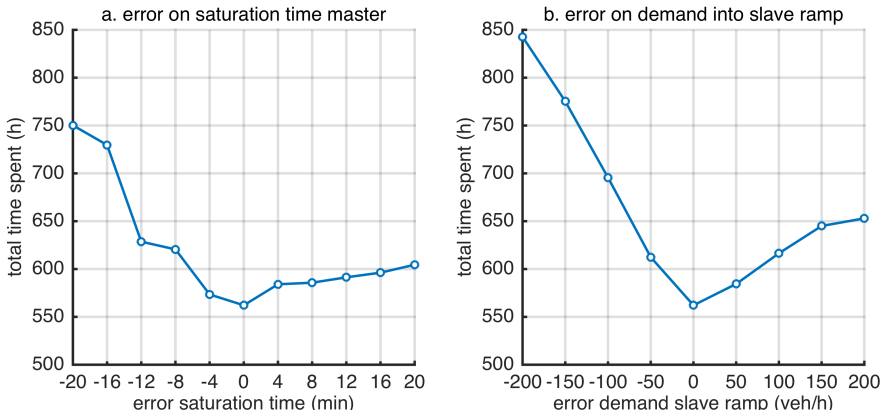

Figure 5.5: Effect on network performance of (a) structural over- and underestimation of the saturation time of the master, and (b) structural over- and underestimation of the demand into the slave ramp.

equalizing available ramp space. Moreover, the strategy is comprehensible, resulting in interpretable coordination metering rates for upstream assisting ramps. However, structural measurement errors on ramp storage spaces and demands into the ramps will result in suboptimal network performance, due to early saturation of either the slave ramps or the master ramp.

Chapter 6

Coordinated intersection control

In this chapter two different (coordinated and integrated) control strategies are proposed to prevent queue spill-back within urban arterials, i.e. to prevent on-ramp queues or intersection queues from spilling back to upstream bifurcation points. The strategies are of the state-feedback type and actively control the network to desired target states given the prevailing network conditions. The controllers can be easily operationalized, they are accurate and able to take unforeseen changes in demand and supply into account. The first strategy fills the allocated storage spaces in parallel, i.e. simultaneously. The second strategy fills allocated storage spaces sequentially, starting with the ones that have the highest fraction of traffic to the bottleneck. In this way, the latter strategy minimizes delay to stored vehicles that do not travel past the bottleneck. The feedback control laws that are responsible for the proper utilization of the allocated storage spaces need tuning. This is a time-consuming task when done based on trial-and-error. A systematic tuning approach is therefore proposed that gives insight into the optimal settings of the feedback gains.

This chapter is based on work published in: Hoogendoorn, S.P., R.L. Landman, J. van Kooten, M. Schreuder, R. Adams, Design and Implementation of Integrated Network Management Methodology in a Regional Network, *Transportation Research Record*, Vol. 2489, pp. 20-28, 2015

6.1 Introduction

As we have seen in the previous chapter, coordinated ramp metering is a means to store vehicles at the ramps located upstream of a freeway bottleneck. The ramp metering duration on such bottleneck can be further extended by storing vehicles at intersections located upstream of the on-ramp. Moreover, as we have seen in Section 4.8.1, it is better to prevent blocking back of queues by temporarily holding back traffic upstream, then to let a queue freely spill back and block an entire intersection.

As we have seen in the state-of-the-art overview of Chapter 2, most of the earlier proposed coordinated intersection control approaches are either not able to properly deal with oversaturated conditions or do not allow for integrated control on an active on-ramp. Optimization based approaches are difficult to employ due to their complexity and high computational demand. From the integrated control approaches that are directly targeting the prevention of a ramp queue spilling back, it is questionable if they are able to properly realize a desired target state, given their rather coarse way of determining control actions.

In this chapter we will therefore propose two different feedback-based control strategies that enable the utilization of urban buffers in real-time to prevent spill-back of queues. The fact that the controllers are feedback-based makes them able to account for unforeseen changes in demand and network characteristics while targeting a desired state. This also allows them to define metering rates that actually stabilize the waiting queues. The first approach aims at filling the allocated storage spaces in *parallel*, meaning that during the buffering process all buffers are filled at the same time. The advantage of this method is that the resulting control law is relatively simple, so that it is easily applied in practice. However, as we have seen in Chapter 4, storing vehicles in all allocated buffers at the same time will cause delay to all traffic within these buffers that does not travel towards the bottleneck. A second approach is therefore introduced that minimizes this hindrance by filling buffers sequentially, one-after the other, starting with the ones that have the highest fraction of traffic to the ramp. Both approaches are of the feedback control type, which means that the involved feedback gains (per allocated buffer) need to be properly tuned to realize stable control behavior. This is,

however, a time consuming task. A tuning approach is therefore designed to systematically determine the optimal setting of the involved gains, so that the resulting system behavior is stable.

By means of simulation tests, the workings of the real-time control approaches are illustrated and their performance compared to each other and to a scenario in which no control is applied. Both approaches strongly improve the network performance, however, applied under the right circumstances, the sequential strategy shows a large improvement over the parallel filling strategy with respect to the delays caused by holding back traffic. The storage space allocation approach that has been introduced in Chapter 4 can again be applied to further minimize the total system delay by choosing the optimal set of buffers.

The structure of the chapter is as follows. In Sections 6.2 and 6.3 we will elaborate on the coordinated control approaches that enable real-time use of urban storage space in both parallel and sequential order. The simulation test case and its results are subsequently discussed in Sections 6.4.The method for tuning the involved feedback gains, in order to realize stable system behavior, is presented in Section 6.5. Finally, the conclusions with respect to the control approaches and tuning method are elaborated in Section 6.6.

6.2 Parallel use of storage space

In this section, the control approach is presented for increasing the effectiveness of ramp metering by deploying multiple intersection controllers upstream of the ramp in the urban arterial. An overview of a typical situation is given in Figure 6.1. The aim of the approach is to saturate all upstream located intersection buffers at the same moment as the on-ramp. In that way all storage space becomes used to extend the metering time on the freeway bottleneck.

6.2.1 Master-slave generalization

The approach is based on the Master-Slave concept proposed in the HERO algorithm for coordinated ramp metering. The general idea is that a Master ramp meter is identified that will start metering on the bottleneck to prevent

or to remove the freeway flow breakdown. In most cases, this will be the first on-ramp upstream of the active freeway bottleneck. Slave on-ramps are subsequently determined that will support the Master in achieving its control task.

This concept can be generalized to include intersection buffers (arms) that are located upstream of the ramp when preventing ramp saturation, or to prevent spill-back of intersection queues over upstream bifurcation points in the urban arterial. To this aim, a master buffer is assigned to meter on a queue that is spilling back, this would normally be a buffer of significant size and with a strongest relation to the bottleneck. Hence, the master can be:

- A single ramp meter (or a metered connection where the ramp meter is supported by the upstream intersection controller) assigned to prevent a freeway flow breakdown;
- An intersection controller that mimics ramp meter functionality to control the inflow to a freeway bottleneck in order to prevent a freeway flow breakdown;
- An intersection controller at the end of an off-ramp that aims at preventing the off-ramp queue from spilling back to the freeway;
- An intersection controller within an urban arterial that prevents spill-back of queues to upstream bifurcation points.

Slaves that assist in the metering task on the considered freeway bottleneck can be the ramp meters upstream of the master (HERO), but they can also be the intersection controllers upstream within the urban arterials. Figure 6.1 shows an example of a ramp meter that is supported by upstream located intersection controllers. The intersection buffers serve as slaves to reduce the inflow to more downstream located links.

A simple feedback mechanism is proposed, allowing these intersection controllers to support the master in achieving its control task as effectively as possible. In doing so, we aim to use the available storage space at the intersection controllers evenly, based on the relative storage space at the master (i.e. in this case the ramp). The algorithm is described in the following section.

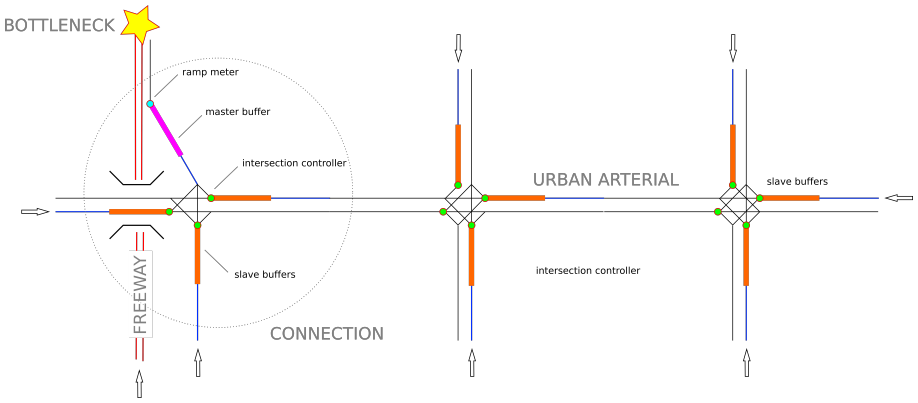

Figure 6.1: An example of a master buffer (the ramp meter) supported by all upstream located intersection controllers. The intersection buffers serve as slaves to reduce the inflow to more downstream located links.

6.2.2 Algorithm design

In order to keep the freeway bottleneck at capacity, the master employs a local ramp metering algorithm. This algorithm determines the metering rate to prevent or remove congestion. ALINEA Papageorgiou et al. (1991) has proven to be a successful approach, and will be adopted in the ensuing. ALINEA determines the ramp's metering rate using a simple feedback algorithm that aims to keep the occupancy at the freeway bottleneck at some (optimal) target value. It uses the on-ramp as the master buffer of which the occupancy will serve as a target value for the slave buffers. The storage space on the ramp is limited to a certain length $s_r^{\max}$ expressed in terms of vehicles. If $w_r(k)$ denotes the current number of vehicles stored on the ramp, then

$$\theta^*(k) = \frac{s_r^{\max} - w_r(k)}{s_r^{\max}} \tag{6.1}$$

denotes the relative storage space left on the on-ramp for discrete time steps k. Note that we assume that all vehicles have the same length (i.e. we work with an average vehicle length). Let B_r denote the set of buffers having a sufficiently strong relation with the ramp, where sufficiently means that the fraction of traffic directed to the ramp is considerable. Ideally the optimal set of applied buffers in the coordination (given their fractions of traffic towards

the bottleneck) is determined using the approach elaborated in Chapter 4. Let $s_b^{\max}$ denote the maximum length of buffers $b \in B_r$ in terms of vehicles. The objective is now to change their outflow such that the relative storage space $\theta_b(k)$ becomes equal to that of the master. To this end, the required buffer outflow $u_b(k)$ is determined by means of a simple feedback strategy

$$u_b(k) = u_b(k-1) + K_b^1 e_b(k) + K_b^2 \Delta e_b(k), \tag{6.2}$$

where $e_b(k) = \theta^*(k) - \theta_b(k)$ and $\Delta e_b(k) = e_b(k) - e_b(k-1)$ for discrete time-steps k. This controller behaves as follows: when $e_b(k) < 0$, the current relative space on buffer b is larger than that of the on-ramp, resulting in a reduced outflow of the considered buffer. Vice versa, if $e_b(k) > 0$ than the considered buffer will release more traffic.

Under ideal circumstances this will result in the situation where all intersection buffers saturate at the same time as the ramp. This in turn implies that all storage space is effectively used to maximally postpone on-ramp saturation and thus the freeway capacity drop. However, there are circumstances that make it impossible to saturate buffers simultaneously, i.e. when the minimum green times make it impossible to hold back enough traffic.

The feedback control law in 6.2 will be applied for each buffer b independently, which obviously does not mean that the buffers are independent (e.g. because the outflow of the one is the inflow of the other). The translation of the metering rate into changes of the intersection control scheme will depend on the scheme at hand. For many Real-Time Network Controllers (such as TOPTRAC or SCATS), metering rates or phase-splits are proposed as boundary conditions of the RTNC to consider during its optimization procedure. For vehicle-actuated controllers, the metering rates are translated into a reduction (or increase) of extension green time.

6.3 Sequential use of storage space

As we have seen in Chapter 4, filling buffers sequentially, starting with those that have the largest fraction of traffic to the bottleneck, reduces the waiting delays. This especially holds when dealing with relatively short peak periods (or disturbances) and when there is a large variation among the coordinated

buffers in their traffic fraction to the bottleneck. However, to ensure that the required inflow reduction to the bottleneck is realized, buffers might need to be grouped into a cluster.

6.3.1 Clustering buffers

The maximum outflow reduction a buffer can effectively realize, is determined by the actual amount of traffic traveling from the buffer towards the bottleneck. When an individual buffer is not able to hold back enough traffic, it needs other buffers to assist in the metering task. Such minimum group of buffers to realize a metering task is identified as a *cluster*. Contrary to the clustering procedure presented in Chapter 4, each buffer can be assigned to a single cluster only. The proposed control algorithm that is introduced in the remainder of this section fills the buffers accordingly. The procedure to define the clusters is given in Figure 6.2 and the corresponding steps 1 and 2 have been discussed in Section 4.8.2.

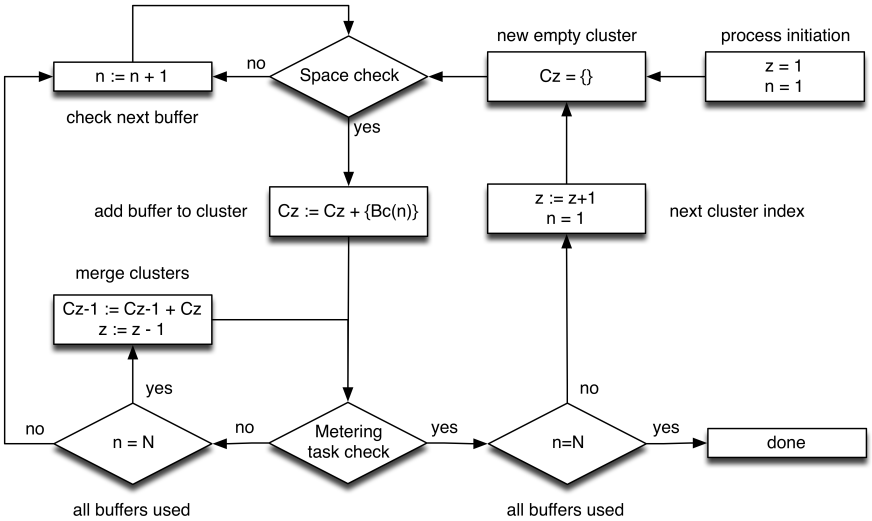

Figure 6.2: Clustering procedure for buffers.

When the clusters of buffers are identified, then the master and slave functions can be assigned. To properly employ the sequential filling strategy in real-time, two additional functions are added for buffers without an active

metering task. Buffers that have been filled are kept full by the Filled function and buffers that have not been used yet are kept empty by the Empty function.

- **Master function.** Within every cluster C_z a master buffer $b = m$ is identified that determines its outflow based on the conditions in the downstream link (saturating ramp or intersection buffer). The master buffer can be chosen based on characteristics that determine its downstream impact such as size, demand and traffic fraction to the bottleneck.

- **Slave function.** In case the metering task cannot be realized by the master buffer alone, assisting (slave) buffers $b \in A_z$ are defined that assist in the metering task. The set of assisting buffers is a subset of the buffers that are assigned to the metering task within the cluster $A_z \subset C_z$. The supportive buffers determine their outflow to the downstream link based on the conditions at the master buffer. The assistance mechanism is then as follows: when the master buffer is saturating, then the assisting buffers start holding back traffic and vice versa.

- **Empty and filled function.** The intersection buffers b that are not part of the active cluster C_z are kept empty in case they are used in a subsequent cluster, and filled in case they have been used in a previous cluster. Keeping a buffer filled effectively means that we allow it to release its incoming traffic demand. Sets are defined for buffers that are kept empty $b \in E_z$ and buffers that are kept full $b \in F_z$.

A conceptual example of a clustering table is presented in Table 6.1, showing the functions of the buffers that are involved in the control process. In total there are three clusters defined:

- Cluster 1 = $\{$buffer 1$\}$;
- Cluster 2 = $\{$buffer 2, buffer 3$\}$;
- Cluster 3 = $\{$buffer 4$\}$.

The boxed functions descriptions mark the buffers that are part of each active cluster. Within a cluster there is always a buffer with the *master* function. In case the cluster consists of multiple buffers, then the other buffers

receive the *assistant* function. Buffers that are not part of the active cluster are kept either *empty* or *filled*. In the remainder of the section we will discuss the dynamic control approach to fill buffers sequentially.

Table 6.1: Functions of the buffers during the different active clusters.

	Buffer 1	**buffer 2**	**buffer 3**	**buffer 4**
Cluster 1	master	empty	empty	empty
Cluster 2	filled	master	assistant	empty
Cluster 3	filled	filled	filled	master

6.3.2 Control loop

In Figure 6.3 the control process is shown. It consists of the elements:

- **Process.** to simulate network states $y(k)$ (in terms of traffic flow, speed and density) based on disturbances $d(k)$ (traffic demands) over time steps k;
- **Monitoring.** to prepare the input data $x(k)$ (relative available storage spaces) for the controllers based on the network state;
- **Control.** to determine the control signals of the intersection $u(k)$ (outflows of the buffers);
- **Actuation.** to operationalize the control signals in terms of effective green times for the involved signal heads;
- **Settings.** to define the setpoints for the applied feedback control laws that control the relative space available within the intersection buffers.

The controller time step is typically larger than the simulation time step of the process model. A clear distinction is therefore made between the simulation time step size T and the time step T_c after which the controller is activated. This results in the time step counters k and k_c denoting time instants kT and $k_c T_c$. We assume T is an integer divisor of T_c:

$$T_c = MT, \tag{6.3}$$

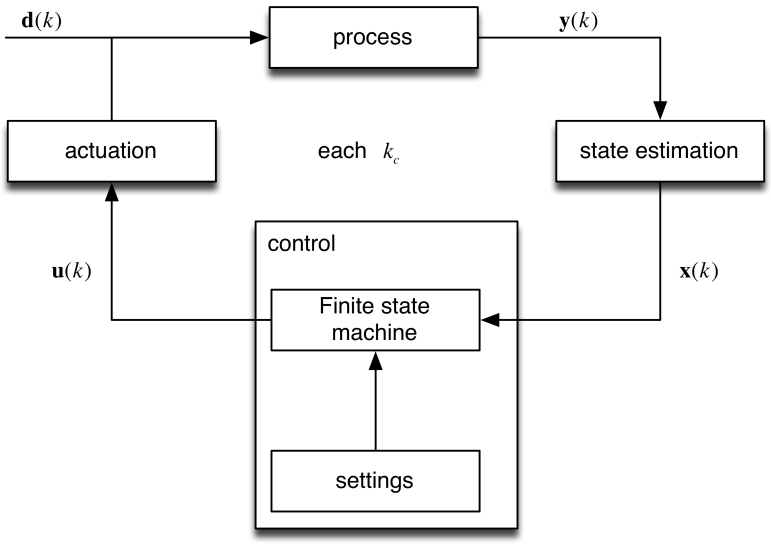

Figure 6.3: Overview of the control loop.

with M an integer. When the controller is activated the corresponding model time index is $k = Mk_c$. The process is modeled by the discrete-time system:

$$x(k+1) = f(x(k), u(k_c), d(k)), \text{if } Mk_c \leq k \leq (k_c+1)M. \tag{6.4}$$

6.3.3 Monitoring

To apply local ramp metering and its integration with upstream intersection controllers, the following information needs to be monitored or determined:

- **Freeway state.** to apply ramp metering on a potential freeway bottleneck. By means of freeway loop detectors the traffic state can be identified in terms of speed, flow and density;
- **Available storage space.** to activate the coordination and determine control signals that sequentially fill the buffers.

The integrated control signals are determined based upon the relative space $\theta_b(k)$ available in each buffer $b \in B$. To this aim, the states of the buffers are monitored in terms of the number of queued vehicles $w_b(k)$. With $s_b^{\max}$ the maximum number of vehicles that can be stored, the relative storage space of a buffer is defined by:

$$\theta_b(k) = \frac{s_b^{\max} - w_b(k)}{s_b^{\max}}. \tag{6.5}$$

6.3.4 Controller design

The control process is governed by the finite-state machine that is presented in Figure 6.4. It consists of three states:

- State 1: Local ramp metering control;
- State 2: Integrated control *with* space at the intersection buffers;
- State 3: Integrated control *without* space at the intersection buffers.

Depending on the prevailing conditions, transitions can be triggered between the various states that are all governed by their own control paradigms. In other words, the finite-state machine functions as a supervisor and decides based on prevailing conditions, when local ramp metering is activated, when the ramp meter will receive assistance, and in which buffers vehicles are delayed given the defined clusters.

The controller needs to properly respond to both oversaturated conditions and undersaturated conditions. During *oversaturated conditions* there is more traffic demand flowing to the bottleneck than its capacity. This will result in queues at the buffers in case control is applied, or in immediate freeway congestion otherwise. During *undersaturated conditions* there is less traffic demand flowing to the bottleneck than the bottleneck has capacity, resulting in the situation where the buffer queues and freeway congestion can dissolve again.

- **Control process during oversaturated conditions.** A cluster is active until its master buffer becomes saturated. Saturation of the master buffer will trigger the activation of the next cluster. The buffers in the previous cluster are then receiving the incentive to remain or become filled, until they are specifically allowed during undersaturated traffic conditions to release stored vehicles.
- **Control process during undersaturated conditions.** During undersaturated conditions, the buffers within the active cluster receive the incentive to release traffic. When the saturation degree of the active

master buffer drops below a certain threshold, then the previous cluster is activated again to also release its stored vehicles. In this way buffers are emptied first of which the space utilization is inefficient due to the low fraction of traffic to the ramp.

It might be difficult to effectively utilize all available buffer space when an assisting buffer has a relatively low demand and large amount of space, because then its state will change slow. The strategy to keep buffers of previously active clusters filled is then especially useful. As assisting buffers from a previous cluster keep receiving the incentive to become filled, they realize an extra flow reduction into the bottleneck. The newly assigned master buffer (from the next cluster) will benefit from this extra reduction. Since the fraction of traffic to the bottleneck of the new cluster is per definition smaller, less traffic needs to be held back with that is not moving past the bottleneck.

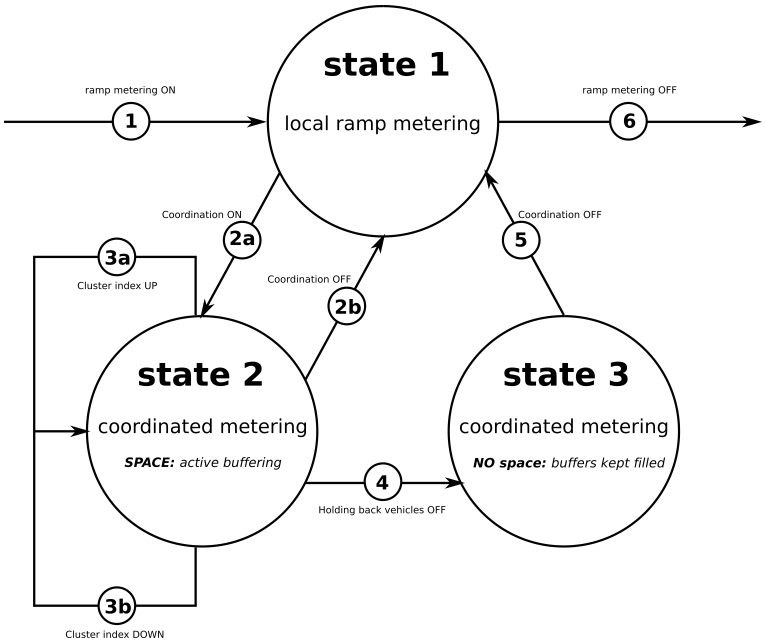

Figure 6.4: Graphical representation of the finite-state machine to activate, deactivate the coordination to support local ramp metering.

State 1: Local ramp metering

As long as the freeway state is functioning well below its capacity, the application of control is not necessary, meaning that the finite-state machine is not active. When the freeway state approximates capacity, the finite-state machine is activated in State 1 where local ramp metering is applied to prevent a flow breakdown in the merging area. Any ramp metering algorithm can be applied that is able to keep the bottleneck at capacity. Activation triggers are algorithm dependent. When applying a feedback based algorithm such as ALINEA, the activation of the local metering mode by *Trigger 1* can be based on the density at a location downstream of the merge

$$\rho_{\mathrm{dn}}(k) > \rho^{\mathrm{loc}}. \tag{6.6}$$

The finite-state machine's output in State 1 then consists of the local ramp metering rate $q_r(k)$ determined by

$$q_r(k) = q_r(k_{\mathrm{c}} - 1) + \alpha(\rho^* - \rho_{\mathrm{dn}}(k)), \tag{6.7}$$

with ρ^* the setpoint value in terms of density corresponding to the capacity downstream the merge, $\rho_{\mathrm{dn}}(k)$ the actual density and α the feedback gain. The control law reduces the inflow to the freeway when the density becomes larger than the desired value to prevent a breakdown. If the density is lower than required, the ramp metering rate is increased to fully use the road capacity.

From the local ramp metering mode in State 1, there are multiple conditions that can activate either a transition to the integrated control mode in State 2 or a complete shutdown of the controller. Let's first have a look at the trigger that activates the integrated control with the upstream intersection to reduce the inflow of the ramp. It is activated at the moment that the relative ramp space $\theta_r(k)$ drops below the threshold θ^{crd}. This is indicated by *Trigger 2a*

$$\theta_r(k) < \theta^{\mathrm{crd}}. \tag{6.8}$$

There are in State 1 also two conditions that can deactivate the ramp metering and the integrated control, indicated by *Trigger 6*. Firstly, there will be no more need to apply ramp metering in case the freeway density downstream of the merging area $\rho_{\mathrm{dn}}(k)$ drops below a certain free flow density threshold ρ^{loc}

$$\rho_{\mathrm{dn}}(k) < \rho^{\mathrm{loc}}. \tag{6.9}$$

Secondly, ramp metering will also be shut down when the ramp queue can be no longer stabilized. We assume that full saturation of the ramp will trigger a shut down of the ramp meter and of the finite-state machine when the ramp runs out of storage space

$$\theta_r(k) < \theta_r^{\mathrm{flush}}. \tag{6.10}$$

There are alternatives to deal with such situation that lead to minor changes in the finite-state machine's design. Instead of a complete shutdown of the ramp meter, a controlled flush metering rate can be used, which allows the ramp queue to dissolve. Whether such adjustments to the design are desired and beneficial depend on the specific conditions at hand.

State 2: Integrated control with space available

Feedback control laws are used to synchronize all involved intersection buffers' actual state with their desired one. The desired states are realized by means of setpoints that correspond to buffers' function in a cluster as illustrated in Table 6.1. State feedback is suitable in this case, because the control delays remain limited due to the small geographical scale of a connection. The target values or setpoints θ^* that are used for the different buffer functions are given in Table 6.2 in terms of relative space.

Table 6.2: Setpoints of the buffers corresponding to the clusters in terms of relative storage space.

Function	Setpoint	Description
Master	θ_r^*	target space to stabilize ramp queue
Assistant	$\theta_{\mathrm{m}}(k)$	actual space at master buffer
Empty	$\theta^{\max}$	target space to keep buffer empty
Filled	$\theta^{\min}$	target space to keep buffer filled

The different clusters that are responsible for stabilizing the ramp queue are activated one after the other. The buffer with the *Master function* determines its outflow based on the conditions at the ramp. When the ramp queue

is longer than desired, the master will reduce its outflow to the ramp and vice versa. This is realized by the following feedback law that controls the signal of the corresponding intersection direction

$$q_b(k) = q_b(k-1) + K_b^1 e_m(k) - K_b^2 \Delta e_b(k), \text{ if } b = m. \tag{6.11}$$

The error between the desired and actual state, and its change over time, are given by $e_m(k) = \theta_r(k) - \theta_r^*$ and $\Delta e_m(k) = e_m(k-1) - e_m(k)$, with $\theta_r(k)$ the current relative ramp space, θ_r^* the target relative ramp space to stabilize the ramp queue, and K_b^1 and K_b^2 the buffer specific feedback gains that need tuning for stable system behavior.

A single feedback law can be used for buffers with an *Assistant, Empty or Filled function*. The only thing that requires adaption is the input for the feedback law, being the setpoint value that corresponds with the buffer's function (see Table 6.2), the buffer's actual state and the buffer's specific feedback gains. The feedback law is given by

$$q_b(k) = q_b(k-1) - K_b^1 e_b(k) + K_b^2 \Delta e_b(k), \text{ if } b \in A_z, b \in E_z, b \in F_z, \tag{6.12}$$

with the error and its change over time given by $e_b(k) = \theta_{\mathrm{b}}(k) - \theta^*$ and $\Delta e_b(k) = e_b(k-1) - e_b(k)$, and K_b^1 and K_b^2 the buffer specific feedback gains.

The buffers $b \in A_z$ assist the master buffer and synchronize their state $\theta_{\mathrm{b}}(k)$ constantly with the state of the master buffer $\theta_{\mathrm{m}}(k)$. When the master buffer is more saturated than the assisting buffer, then the assisting buffer reduces its outflow and vice versa. This mechanism ensures that the master buffer and its assistant buffers collaborate such that the aggregated inflow to the ramp becomes equal to its outflow, while all buffers have the same saturation degree and be filled at the same time.

The buffers $b \in F_z$ that have been filled in previous clusters are kept full and buffers $b \in E_z$ that are allocated to upcoming clusters are kept empty. To this aim, the buffers' actual state $\theta_b(k)$ is synchronized with setpoints $\theta^* = \{\theta^{\max}, \theta^{\min}\}$ respectively, indicating the desired maximum and minimum fill.

In State 2 of the finite-state machine there are four different conditions that either trigger a cluster change (within State 2) or a transition to another control state. Lets first review the triggers that are activated during *oversaturated conditions*. When the master buffer $b = m$ of an active cluster C_z with $z < Z$ becomes saturated, the next cluster $z := z + 1$ is activated by *Trigger 3a*

$$\theta_{\mathrm{m}}(k) < \theta^{\min} + \varepsilon \ \wedge \ \Delta\theta_m(k) > 0, \tag{6.13}$$

where $\Delta\theta_m(k) = \theta_m(k-1) - \theta_m(k)$ indicates if the master is running out of space. The term ε indicates an extra threshold value to prevent the finite-state machine from oscillating between clusters. At the moment the master buffer of the last defined cluster $z = Z$ is filled, a similar *Trigger 4* is used to activate a transition from State 2 to State 3 where all involved intersection buffers are kept full

$$\theta_{\mathrm{m}}(k) < \theta^{\min} + \varepsilon \ \wedge \ \Delta\theta_m(k) > 0. \tag{6.14}$$

When cluster index $z = Z$, there are no more buffers with storage space available and in the description of State 3 it is explained why the buffers are then kept full.

When the freeway bottleneck becomes *undersaturated* before all available storage space has been used, then both the ramp and the buffers are allowed to release their stored vehicles. At the moment that the active master buffer is emptied, then *Trigger 3b* makes the finite-state machine shift the cluster index down $z := z - 1$ to empty the buffers of the previous cluster

$$\theta_m(k) > \theta^{\max} - \varepsilon \ \wedge \ \Delta\theta_m(k) < 0, \tag{6.15}$$

with $\Delta\theta_m(k) = \theta_m(k-1) - \theta_m(k)$. When all stored vehicles are released from the intersection buffers, the ramp queue will also fully dissolve. The moment that the actual ramp space $\theta_r(k)$ drops below the threshold θ_r^{crd} the system will switch back to the local ramp metering mode in State 1. This is indicated by *Trigger 2b*

$$\theta_r(k) > \theta^{\mathrm{crd}} + \varepsilon. \tag{6.16}$$

State 3: Integrated control without space available

Keeping all involved buffers filled at a predefined target value in State 3, implies that the outflow of a buffers becomes equal to its incoming demand. The active feedback law is given in 6.12 in combination with the setpoint $\theta^{\min}$ from Table 6.2.

There are two conditions that trigger a state transition to the local metering mode in State 1. During oversaturated conditions, keeping the buffers filled ensures that the storage space at the intersection buffers and the storage space at the ramp become fully used. By keeping the queues in the intersection buffers at constant length, the outflow to the ramp is increased. This results in the ramp's full saturation. If during State 3, the conditions become undersaturated, then keeping buffer queues at constant length results in a decreasing ramp queue. Both situations are captured by *Trigger 5* activating a transition from State 3 to State 1. During oversaturated conditions the trigger is

$$\theta_r(k) < \theta^{\min} - \varepsilon, \tag{6.17}$$

and during undersaturated conditions

$$\theta_r(k) > \theta^{\text{crd}} + \varepsilon. \tag{6.18}$$

Back in State 1, the finite-state machine decides to either shut down the ramp metering (if the ramp remains saturated) or to continue local ramp metering (if the ramp has space available). In the latter case, the finite-state machine might again become activated, or the freeway conditions might allow for a shut down of the ramp metering installation.

6.4 Test case

In this section we will evaluate by means of a simulation test case, the workings and performance of the proposed integrated control concepts for using urban storage space to extend the ramp metering time on a freeway bottleneck. The first controller employs the parallel filling strategy, meaning that the applied storage spaces are filled simultaneously and the second controller

realizes a sequential filling strategy such that buffers are filled in the order of decreasing traffic fraction to the bottleneck.

The applied traffic network and its characteristics are given in Figure 6.5 and Table 6.3. The network consists of a freeway stretch with an on-ramp along side where urban traffic can access the mainstream. The intersection controller upstream of the ramp consists of three arms, where traffic can be temporarily stored. To this aim, the green time of the corresponding signal (see magenta numbering for signal identities) are shortened. Traffic fractions are used to define the amount of traffic that moves from the buffer to the freeway. These fractions influence the effectiveness with which the storage space is used, and hence, the duration with which the metering duration of the ramp can be extended. In combination with the total demand, the fractions determine the maximum outflow reduction that a buffer is able to realize.

Table 6.3: Ramp and intersection buffer characteristics: storage space in terms of meters, the number of lanes, the turn fraction of traffic to the ramp, and the min. and max. outflow of the signals to the ramp (i.e. signals 1, 5 and 9).

storage place	Length (m)	Lanes (-)	Turn fraction (%)	Min. outflow (veh/h)	Max. outflow (veh/h)
Ramp	500	2	1	100	4000
Arm 2	500	2	1	100	600
Arm 5	500	2	0.5	100	600
Arm 8	500	2	0.4	100	600

With respect to the sequential filling strategy, a 3-cluster and a 2-cluster setup are presented. In the first setup each cluster consists of a single buffer to execute the metering task. In the other setup, the second cluster consists of a master buffer that is directly metering on the ramp and an assisting buffer that supports in the metering task. The network performance of the 3-cluster sequential strategy is then compared to the parallel strategy to illustrate its benefits.

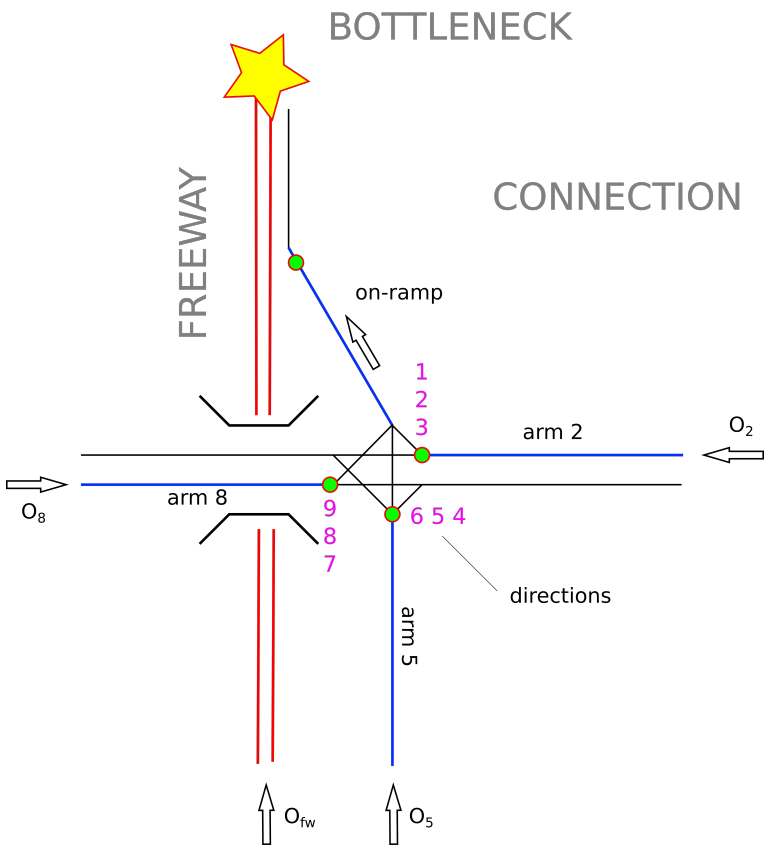

Figure 6.5: Network layout test case consisting of a freeway with one connection along side. The ramp is fed by three intersection arms, respectively arms 2, 5 and 8, identifying the corresponding buffer identities. All three buffers are included in the coordination to reduce the inflow to the ramp.

During the simulation a bottleneck is created by making the total flow towards the downstream end of the freeway larger than capacity. The peak period subsequently determines the duration of the oversaturated phase, after which the peak demand is instantly lowered to an off-peak demand. During the oversaturated phase, ramp metering keeps the bottleneck at capacity as long as there is space available at the ramp and intersection buffers. A queue detector is located at the upstream end of the ramp and the intersection buffers. Ramp metering and coordination are shut down once the detectors are triggered. This queue protection mechanism instantly releases stored vehicles into the mainstream, causing a flow breakdown in the bottleneck.

Traffic is generated at origin O_{fw} on the freeway and at origins O_2, O_5 and O_5 at the intersection buffers. The fractions γ_b, respectively 1, 0.5 and 0.4 (see Table 6.3) determine the flow towards destination D_{fw} downstream the freeway. The demands over time are given in Table 6.4 for a peak period of one hour, clearly distinguishing a peak and off-peak flow. During the simulation the demands are defined per 15 minutes, which makes the change from peak to off-peak demand happen within a 15 minute interval.

Table 6.4: Traffic demand loaded at the origins in terms of veh/h.

Time (hh:mm)	7:00	7:30	8:00	8:30	9:00	9:30
O_{fw}	3000	3000	3000	1500	1500	1500
O_2	400	400	400	200	200	200
O_5	800	800	800	400	400	400
O_8	1200	1200	1200	600	600	600

The difference in system performance is caused by the intersection buffer outflows, given the strategy with which they are filled. The system performance is determined by the total network outflow. The metering duration of the ramp is independent of the buffer filling strategy, since all space becomes fully used to prevent ramp saturation.

6.4.1 Applied traffic flow model

The macroscopic multi-class cell-based traffic flow model Fastlane van Lint et al. (2008) has been used for the process simulation. Fastlane is able to reproduce the phenomena that influence the network performance, i.e. the build up and dissolving of congestion, blocking back effects of queues and the capacity drop. Fastlane applies a Godunov-scheme to model the flows between cells. In case a cell becomes congested, the supply of directly downstream located cells is decreased by a certain factor to reproduce the capacity drop (e.g. 15%).

6.4.2 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 all vehicles in the network (incl. queues on the on-ramps) over a period $k = \{0, 1, ..., K-1\}$ with K the total number of simulation time steps is determined by:

$$J_{\mathrm{TTS}} = T \sum_{k=1}^{K} \sum_{m \in M} \sum_{c \in C_m} \rho_{m,c}(k) \lambda_{m,c}, \tag{6.19}$$

with $\rho_{m,c}(k)$ the vehicle densities (veh/km) over the cells $c \in C$ of all links in the network $m \in M$ and $\lambda_{m,c}$ the corresponding cell lengths (in km). The total delay (TD) for each scenario is determined by subtracting the TTS during free flow conditions from the TTS of each control scenario.

6.4.3 Setup of the algorithms

The parallel and sequential filling strategy are executed by the finite-state machine presented in Section 6.3.4. It determines when the coordination is activated (i.e. State $1 \rightarrow 2$) and deactivated (i.e. State $2 \rightarrow$ State 1). However, State 2 for the parallel filling strategy makes use of a control law that does not involve the use of multiple clusters. The following settings apply for both strategies:

- **Activation coordination.** The activation threshold (Trigger 1) is chosen $s^{\mathrm{crd}} = 20\%$, meaning that traffic will be held back at the intersection as soon as the available space at the ramp drops below 20%;

- **Deactivation coordination.** The threshold in terms of relative ramp space that terminates the coordination (Trigger 2) and move the controller back to local metering mode is $s^{\text{loc}} = 50\%$;

- **Flush stored traffic.** The threshold to activate a flush of stored traffic at the intersection buffers or the ramp queue is chosen $s^{\text{flush}} = 5\%$ in terms of relative storage space. For the integrated intersection buffers this only applies to the buffers that are actively used. There is no need to apply this threshold for a buffer that is kept full during the sequential filling strategy, because the controller will maximize the outflow if needed;

- **Feedback gains.** The feedback gains are tuned based on the approach elaborated in Section 6.5, resulting in $K_b^1 = 300$ and $K_b^2 = 3500$ for all coordinated buffers $b \in B^{\text{c}}$.

For the parallel filling strategy the relative storage space at all buffers $s_b^{\text{rel}}(k)$ is synchronized with that of the ramp $s_r^{\text{rel}}(k)$. In this way, the buffers saturate at the same time as the ramp. This requires no further setup of parameters. The setup of the sequential filling strategy is a bit more involved, because we also need to set target values for:

- **Stabilizing ramp queue.** The target value in terms of relative storage space for the keeping the queue at the ramp at constant length is chosen $s_{\text{r}}^* = 80\%$;

- **Keeping buffers full or empty.** The target values in terms of relative storage space for keeping buffers filled and empty are respectively $s^{\min} = 10\%$ and $s^{\max} = 90\%$;

- **Switching clusters.** the threshold for switching the active cluster up and down in terms of relative storages space are chosen $\varepsilon = 10\%$, resulting in trigger 3 $s^{\min} + \varepsilon$ and trigger 4 $s^{\max} - \varepsilon$.

6.4.4 Test case results

In this section the results are presented. First the parallel filling strategy and then the 3- and 2-cluster setups for the sequential filling strategy are discussed.

Parallel filling strategy

The ramp and intersection buffers saturate at approximately the same moment, if the demand and control characteristics allow for this to happen. This control strategy can be seen in Figure 6.6.

- Integrated control is activated around 7:20h at the moment that the relative ramp space drops below 20%;
- All intersection buffers start holding back traffic to postpone the on-ramp saturation;
- At first the ramp storage space increases and later it decreases parallel to the relative storage spaces of the intersection buffers until all space has run out;
- At 8:15h the queue detector is activated, triggering the release of stored traffic from the intersection buffers;
- The ramp saturates, leading to a flush of the ramp queue and a freeway flow breakdown.

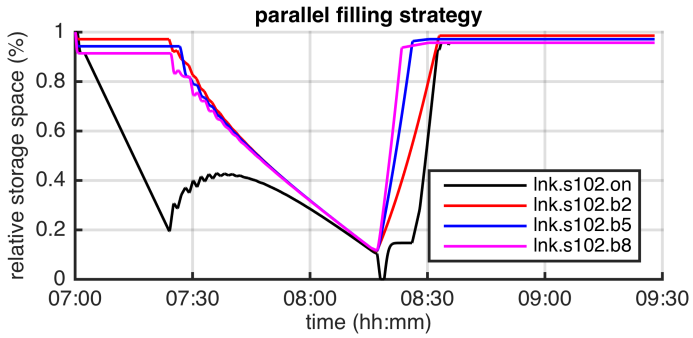

Figure 6.6: The relative space at the ramp and buffers for the parallel filling strategy. In black the ramp storage space, and in red, magenta and blue lines the storage space at respectively intersection buffers 2, 5 and 8.

Sequential filling strategy: three clusters

If there is a large variance in buffers' turn fraction to the ramp, it is best to apply a sequential filling strategy to reduce storage delay. Within this test case, all buffers are able to individually realize the metering task on the ramp. Hence, the network performance is improved by sequentially activating each buffer. The strategy can be seen in Figure 6.7a:

- The coordination becomes activated around 7:20h as the ramp's storage space drops below the 20%. The target value for the ramp space is 10%, which means that the intersection buffers do not have to hold back traffic yet;

- At 7:30h the master buffer of the first cluster starts decreasing its outflow to keep the ramp queue at constant length (see Figure6.7b). Since its turn fraction is 100%, the rate with which its space depletes is similar to that of the ramp;

- Just before 8:00h this buffer's storage space drops below 20% and then the next cluster is activated. The new master buffer (blue line in Figure 6.7a and b) does not need to store traffic yet, since the previous master buffer is still aiming at becoming completely filled at a storage space of 10%(see Figure 6.7d). To reach this target value the outflow remains lower than the demand for some more time, which can be considered assistance to the next master;

- At 8:00h the storage space of this new master starts decreasing with a higher rate than that of the ramp and previous master (see Figure 6.7a). The reason is the lower fraction of traffic that turns to the ramp, meaning that more traffic is held back to realize the desired ramp inflow reduction. The fact that storage space becomes less efficiently used can also be seen by the shorter time intervals with which a cluster is active;

- The final cluster saturates around 8:15h, leading to a release of all stored traffic which is associated with a freeway flow breakdown;

- Buffers are kept empty that have not become active yet within previous clusters (see Figure 6.7e). This is desirable, given the fact that such buffers are causing more hindrance to ongoing vehicles that are not passing the ramp.

Sequential filling strategy: two clusters

The sequential filling strategy with two clusters is very similar to the workings of that with three clusters. However, the second cluster consists of a master buffer and an assisting buffer that together realize the metering task on the ramp. Figure 6.8a nicely shows how both buffers are simultaneously filled up to the point where all storage space runs out. Figure 6.8a shows the corresponding master functions of buffers 2 and 5 and Figure 6.8a identifies the slave function of buffer 8. In Figure 6.8d it can be seen how buffer 2 is kept full as cluster 2 is activated, and Figure 6.8e shows how the buffers lose their empty function once activated.

Benefits of the sequential over the parallel strategy

In this section the sequential and parallel filling strategy are compared in terms of total delay. To determine the delay for each controlled scenario, the total time spent under free flow conditions is subtracted from the total time spent as a result of the application of integrated control. In Figure 6.9a the total time spent curves are shown as a function of the peak period. The longer the peak period lasts, the more vehicles travel within a scenario. It is clearly visible that the curve of the parallel strategy is always higher than that of the sequential strategy. The question remains to what extent the delays are reduced by the sequential strategy. The relative improvement in terms of total delay is given in Figure 6.9, showing a maximum improvement of almost 25% in this case for a peak period of half an hour. To conclude, as can be seen in Figure 6.9c the benefits in terms of total time spent of the parallel and sequential strategy over the no-control case can increase up to respectively 12% and 15%.

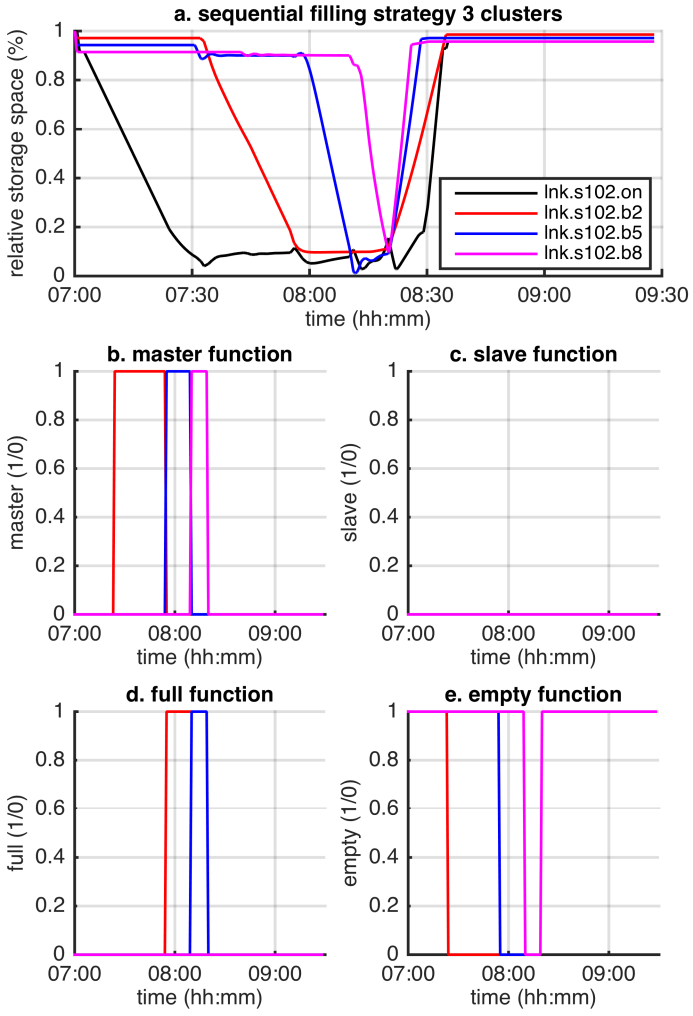

Figure 6.7: The relative space at the ramp and buffers for the 3-cluster sequential filling strategy, with (a) in black the ramp storage space, and in red, magenta and blue the storage space at intersection buffers 2, 5 and 8, (b) the master function, (c) the slave function, (d) the full function and (e) the empty function.

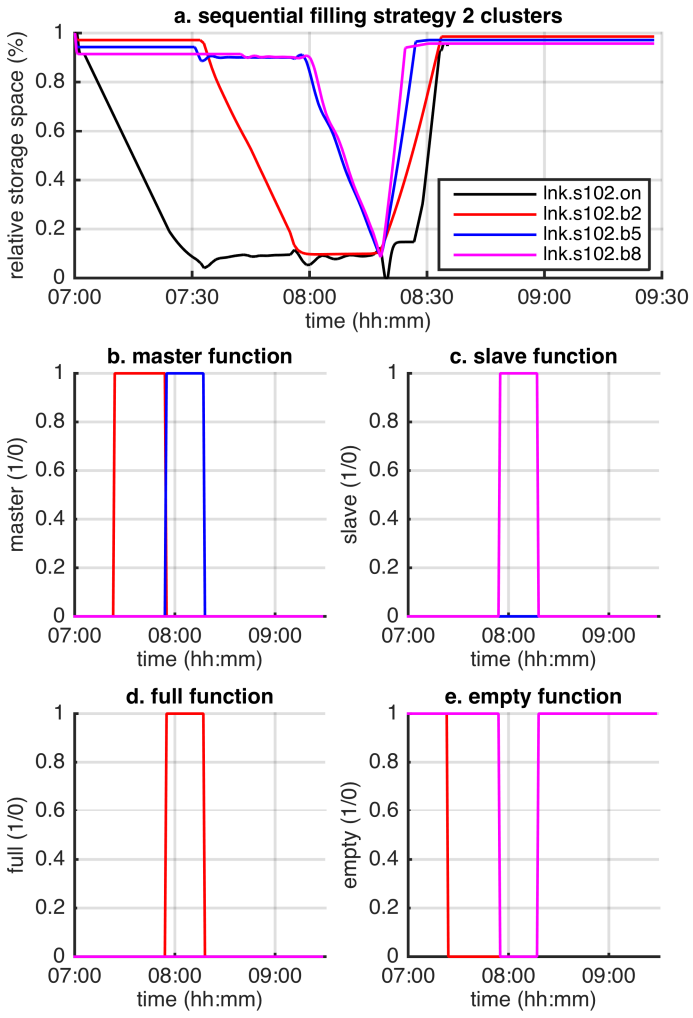

Figure 6.8: The relative space at the ramp and buffers for the 2-cluster sequential filling strategy, with (a) in black the ramp storage space, and in red, magenta and blue the storage space at respectively intersection buffers 2, 5 and 8, (b) the master function, (c) the slave function, (d) the full function and (e) the empty function.

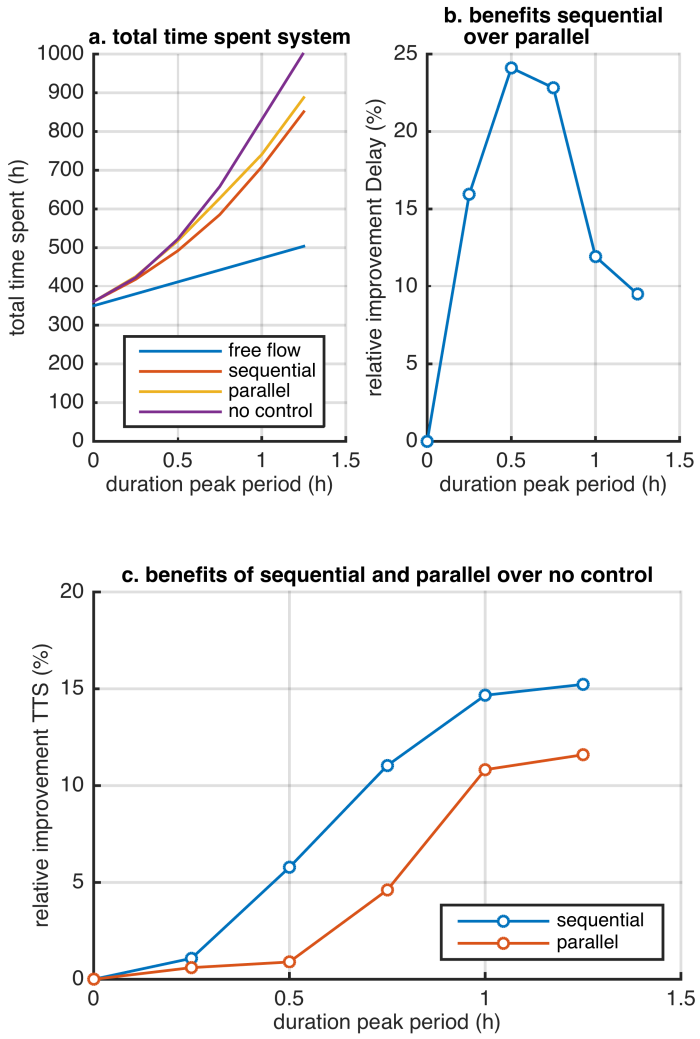

Figure 6.9: Comparison of the parallel and 3-cluster sequential filling strategy, with (a) the network performance in terms of total time spent, (b) the relative improvement of the sequential over the parallel strategy in terms of delay, and (c) the relative improvement of the sequential and parallel strategy over the no-control case in terms of total time spent.

6.5 Tuning approach for feedback gains

Tuning the gains of state feedback controllers is often done based on a trial-and-error procedure, which can be a very time consuming task. The gains that are applied in a state feedback control law determine to what extent the previous control signal is adjusted based on the difference between the target state and the current state estimate and the evolution of this error over time. Choosing the gains too large would result in an aggressive controller that makes large adjustment in the control signal to achieve some desired target state. If the gains are chosen too small, then the controller adjusts its signal too slow to adequately respond to changing conditions.

Proper tuning of the gains is dependent on the control interval that determines how often the signal is updated and the system dynamics that determine how fast conditions are changing. Even with a good understanding of the control process, it would still be a time consuming task to tune the parameters based on trial-and-error procedure. Moreover, the parameters require retuning when network, demand or control characteristics change.

In this section, a method is therefore put forward that enables mathematical analysis of the controller and optimal choice of the involved feedback gains. It helps finding optimal gains that reduce the error as fast as possible given typical conditions. Once the approach is setup for the situation at hand, then new feedback gains can be determined in no-time if conditions change.

The methodology is elaborated based on the situation where the inflow of an on-ramp is controlled by means of a single intersection buffer. Once the method is fully elaborated and its application illustrated by means of a worked example, then two more complex cases are discussed. The introduced methodology is extended for different situations that involve more intersection buffers.

It takes time for a vehicle to influence the state of the downstream storage space. Hence, when the impact of control signals is delayed, feedback gains need to be tuned such that system stability remains guaranteed. The impact

of the intersection buffer outflow on the ramp queue is therefore modeled by means of system variables.

6.5.1 System description

In this section we will first model the system dynamics and cast it in a state space definition. This enables a stability analysis on the resulting system (i.e. including the active controller). The method assumes a fixed metering rate of the on-ramp on a freeway bottleneck. The aim of the control law is then to synchronize the conditions at the upstream buffers as fast as possible with that of the ramp. Finding the optimal settings of the feedback gains will result in the situation where the ramp and buffer become saturated at approximately the same moment.

On-ramp with a single buffer

First the mathematical model is derived for the situation where the ramp state is influenced by the outflow of a single buffer as given in Figure 6.10. Let $s_r(t)$ denote the current storage space on the on-ramp in terms of vehicles and its dynamics is given by

$$s_r(t+1) = s_r(t) + (q_r(t) - d_r(t) - \alpha u(t-T))\Delta t, \tag{6.20}$$

with $q_r(t)$ denotes the ramp-metering rate expressed in vehicles per hour, which is assumed to be determined by the ramp-metering strategy and thus given; d_r denotes the (uncontrolled) autonomous ramp inflow (e.g. from non-controlled buffers); Δt denotes the time step. The signal $u(t)$ is the controlled outflow from the buffer; T denotes the time-delay caused by the travel time between the buffer and the on-ramp (excluding the additional delay due to queuing). Finally, α describes the fact that not all traffic from the buffer flows into the ramp. The buffer space $s_b(t)$ on the buffer then satisfies the following equation

$$s_b(t+1) = s_b(t) + (u(t) - d_b(t))\Delta t, \tag{6.21}$$

where $d_b(t)$ denotes the autonomous traffic demand into the buffer. In order to prevent the ramp from saturating, the aim is to synchronize the relative space of the buffer with that of the ramp. Storage space in terms of vehicles such as $s_b(t)$ can be easily translated in to relative storage space $\theta_b(t)$ by:

$$\theta_b(t) = \frac{s_b^{\max} - s_b(t)}{s_b^{\max}}, \tag{6.22}$$

with $s_b^{\max}$ the maximum storage space in terms of vehicles. In line with the parallel filling approach proposed in Section 6.2, the buffer metering rate $u(t)$ is determined by a feedback law that aims at synchronizing the buffer's state in terms of relative storage space with that of the ramp

$$u(t+1) = u(t) + K^1 e(t) + K^2 \Delta e, \tag{6.23}$$

with $e_b(t) = \theta_r(t) - \theta_b(t))$ and $\Delta e_b(t) = e_b(t) - e_b(t-1)$. As can be seen from the above equations, the storage space dynamics are now expressed as a function of demand input and the control law for storing vehicles upstream of a critical link. Our aim is subsequently to analyze the impact of the control gains K^1 and K^2 on the system behavior.

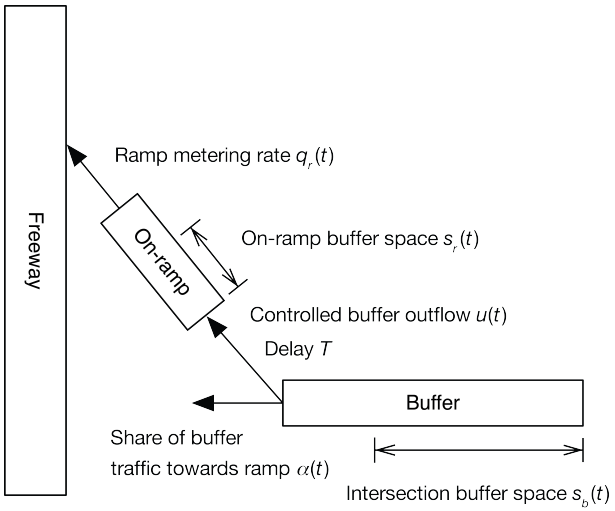

Figure 6.10: Overview of the network situation and corresponding variables where the ramp state is influenced by the outflow of a single upstream intersection buffer.

6.5.2 Discrete time state-space formulation

To this end, we recast our model in a simple discrete-time state space formulation. We define the state vector $\vec{x}(t)$ by considering all information in the system that is needed to predict the state at time instant $t+1$. There are two issues that are somewhat complicating matters: the fact that the model has an explicit delay T and the fact that the controller uses differences of the error (i.e. $\Delta e(t) = e(t) - e(t-1)$). To get rid of the explicit delays, the state is defined as follows:

$$\vec{x}(t) = \{s_r(t), s_b(t), u(t), s_r(t-1), s_b(t-1), u(t-1), ..., u(t-T)\}. \quad (6.24)$$

Using this definition, the dynamics of our model can be written as a state-space equation, i.e. $\vec{x}(t+1) = \vec{f}(t, \vec{x}(t), \vec{d}(t))$. More specifically, the model can be cast in a linear form:

$$\vec{x}(t+1) = A\vec{x}(t) + \vec{b}(t), \quad (6.25)$$

with

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & \cdot & -\alpha\Delta t \\ 0 & 1 & \Delta t & 0 & 0 & 0 & \cdots & 0 \\ \frac{K_1+K_2}{L_r} & -\frac{K_1+K_2}{L_b} & 1 & -\frac{K_2}{L_r} & \frac{K_2}{L_b} & 0 & \cdots & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \ddots & 0 \end{bmatrix} \quad (6.26)$$

and

$$\vec{b}(t) = \begin{bmatrix} q_r(t) - d_r(t) \\ -d_b(t) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \Delta t. \quad (6.27)$$

In other words, the prediction of the next state becomes a matrix multiplication of the previous one.

6.5.3 Stability analysis

In order for the system to be stable it needs to move towards some steady state, which in this case implies that the control error defined by the state variables $s_r(t)$ and $(s_b(t))$ needs to become zero (as fast as possible). To evaluate this system behavior, matrix A can be decomposed as the product of two arbitrary matrices Σ and Σ^{-1} and Λ a diagonal matrix with the eigenvalues of A (corresponding to the elements $x_i \in \vec{x}(t)$) on its diagonal:

$$A = \Sigma\Lambda\Sigma^{-1} = I\Lambda, \tag{6.28}$$

which in turn over N prediction time steps would lead to

$$A^N = \Sigma\Lambda^N\Sigma^{-1} = I\Lambda^N. \tag{6.29}$$

For each of the elements $x_i \in \vec{x}(t)$ stability can be assured if the the absolute value of the corresponding eigenvalue λ_i is smaller or equal to 1, i.e. the corresponding variable will move to a steady state. In case $|\lambda_i| < 1$, the rate at which $x_i(t)$ converges is determined by how small the value of the eigenvalue is (smaller means faster convergence). Hence, the faster the variables converge that describe the system error, the faster the controller reaches its target value or set point.

Only the variables describing the control error are of interest, i.e. the main objective of the controller is to reduce the error $e(t) = x_1(t)/L_r - x_2(t)/L_b$ as quickly as possible. Defining the vector $\vec{k} = \{1/L_r, -1/L_b, 0, ..., 0\}$, resulting in $e(t) = \vec{k}'\vec{x}(t)$, indicates which elements of $\vec{x}(t)$ actually contribute to the error $e(t)$ which is aimed to be minimized. Only these eigenvalues λ_i are therefore considered in the stability analysis. In sum, the maximum relevant eigenvalues determining the dynamics of the controller error will be considered:

$$\gamma = \max\{|\lambda_i| \text{ subject to } \{\vec{k}'\vec{x}(t)\} \neq 0\}. \tag{6.30}$$

6.5.4 Worked example

The methodology is illustrated for a system with a 2-minute delay ($T = 2s$) and a system with a 5-minute delay ($T = 5s$). The remaining parameters are set as follows: $\alpha = 1$, $L_r = 100m$, $L_b = 50m$ and $\Delta t = 1min$. Figure 6.11 shows the stability diagrams for both cases, showing the maximum relevant

value of $|\lambda_i|$. From the figure it can be seen under which choices for K_1 and K_2 the controller is stable or even has the most rapid reduction of the error $e(t)$ (i.e. the lowest values of the maximum eigenvalue). The contours are quite different, but still choices can be made that would yield reasonable behavior for both controllers. Also note that $K_2 > 0$ is necessary to stabilize the controller (eigenvalues are always larger than 1 for $K_2 = 0$).

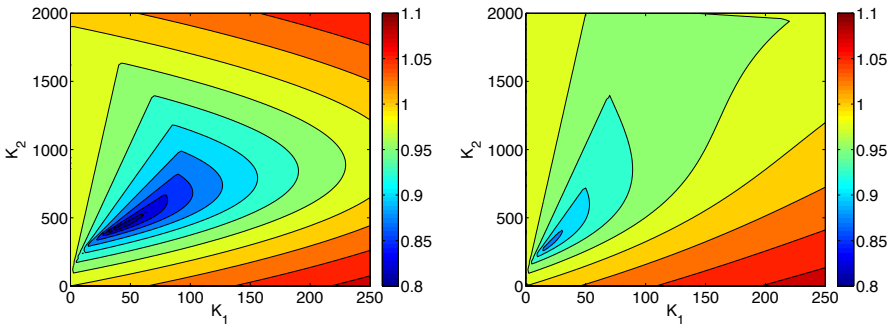

Figure 6.11: Stability diagrams showing γ as a function of the control gains, for $T = 2$ (left) and $T = 5$ (right).

To illustrate what it means for the controller behavior, we consider three scenarios (choices for the gains) in different stability regions. Figure 6.12 shows the results of this analysis, comparing an optimal situation $(K_1, K_2) = (42.5, 440)$, a near optimal situation $(K_1, K_2) = (82.5, 440, 440)$ and an unstable situation $(K_1, K_2) = (225, 1500)$. As can be seen by the blue continuous line, an optimal choice of gains quickly reduces the difference (error) between the state at the ramp and the state at the buffer to zero. Gains from a the near optimal region also manage to remove the system error, however, the fluctuation in error is larger and it takes longer to realize a stable state. Choosing gains from the unstable region makes it impossible for the controller to synchronize the state of the buffers with that of the ramp, as can be seen from the red line showing how the error amplifies over time.

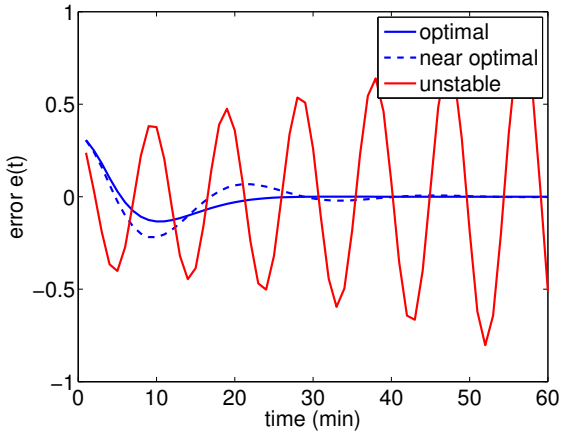

Figure 6.12: Error dynamics for combinations of feedback gains from the optimal, near optimal and unstable regions.

6.5.5 Extension of the methodology

In this section the method is extended for the use of buffers that are used series and in parallel. In the first case the upstream located outflow of a buffer influences the state at a downstream buffer and in the latter case multiple buffers influence the inflow of the downstream located ramp.

Controlling serial buffers

The second situation that will be considered is a case where another buffer connects to the buffer feeding the on-ramp. In other words, the flow into the buffer is the (controlled) outflow of the second buffer. This gives us for the buffer leading to the ramp:

$$s_b^1(t) = s_b^1(t) + (u_1(t) - u_2(t-T))\Delta t, \tag{6.31}$$

while for the second buffer we have:

$$s_b^2(t) = s_b^2(t) + (u_2(t) - d_b^2(t-T))\Delta t. \tag{6.32}$$

For either buffer, the controlled outflow is given by 6.23, where the gains for each buffer can be set separately. The state $\vec{x}(t)$ can be defined as done before and the system can then be written in the discrete-time state space form.

Again, we can establish the dynamics of the systems and in particular the two errors $e_1(t) = s_r(t)/L_r - s_b^1/L_b$ and $e_2(t) = s_r(t)/L_r - s_b^2/L_b$ by studying the eigenvalues of A.

The worked examples for this situation consider a delay $T_1 = T_2 = 2s$ and the length of both buffers is chosen $L_b^1 = L_b^2 = 50m$. All other parameters are similar to the first case. Figure 6.13 shows the maximum absolute eigenvalues for different choices of the gain parameters of the controller, clearly showing that in particular the choice for good parameters for buffer 2 is important for the stability of the system.

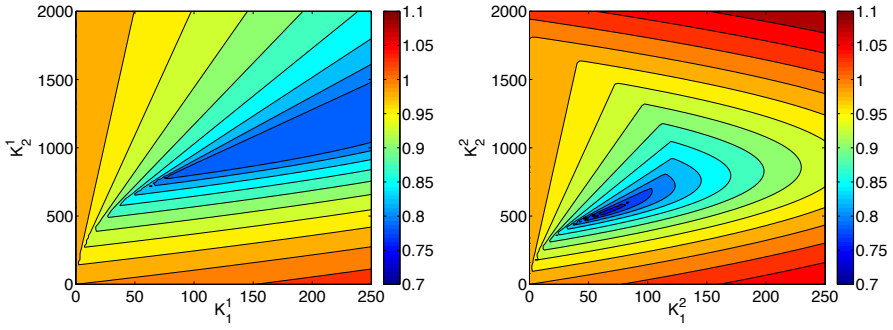

Figure 6.13: Maximum eigenvalues γ sequential buffers for both changes in the gains of a) buffer 1, assuming a near optimal choice for K_1^2 and K_2^2 and of b) buffer 2, assuming a near optimal choice for K_1^1 and K_2^1.

Controlling parallel buffers

The last example that we will consider is the situation in which two intersection buffers feed into the ramp. In this case, the on-ramp storage space is modeled as follows:

$$s_r(t+1) = s_r(t) + (q_r(t) - d_r(t) - \alpha_1 u_1(t - T_1) - \alpha_2 u_2(t - T_2))\Delta t. \quad (6.33)$$

The equations describing the dynamics of the intersection buffers are in line with the first example. We have tested the controller stability in the case where $T_1 = 5s$ and $T_2 = 2s$. Figure 6.14 shows the stability analysis re-

sults. First of all, the graphs show that the controller is stable for the correct choices of the controller gains, even if the details of both buffers are not the same. Depending on the delays, the choices are quite different.

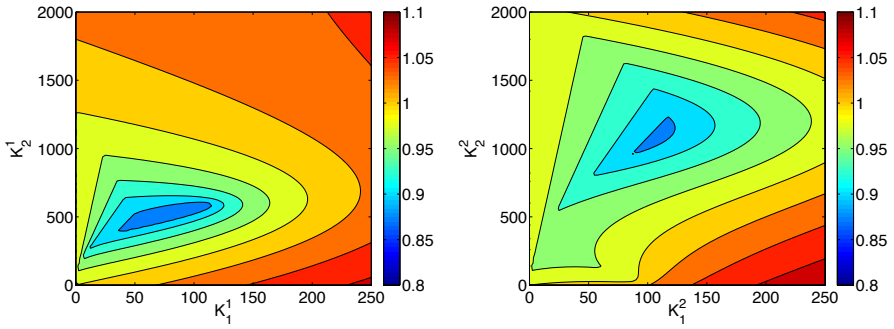

Figure 6.14: Maximum eigenvalues γ parallel buffers for both changes in the gains of a) buffer 1, assuming a near optimal choice for K_1^2 and K_2^2 and of b) buffer 2, assuming a near optimal choice for K_1^1 and K_2^1.

6.5.6 Conclusions tuning approach

In this section, an approach has been presented that provides insight into the characteristics of the control algorithm, in particular looking at the stability of the controller. The presented method allows choosing the controller gains such that the rate at which the controller error goes towards zero is maximized, without having to resort to trial-and-error. This systematic approach saves a considerable amount of time in tuning the gains and in adjusting them when conditions change.

6.6 Conclusions and discussion

In this chapter we have proposed two methods for holding back traffic within the urban network to postpone the spill-back of ramp and intersection queues. When spill-back of the ramp queue is controled, the capacity drop phenomenon is longer prevented. The first approach is a parallel filling strategy, that simultaneously fills allocated storage spaces based on the available

space at the critical bottleneck link. The disadvantage of this method is that ongoing vehicles (i.e. that have nothing to do with the bottleneck) become hindered in all buffers from the moment the coordination is activated. In order to minimize this hindrance, the second approach sequentially fills buffers in decreasing order with respect to the fraction of traffic to the bottleneck. An important prerequisite is that clusters of buffers need to be generated that each can hold back sufficient traffic to stabilize a queue. The test case results show that large benefits in terms of delay reduction can be achieved (e.g. 25% in the presented test case), if there are large variations in buffers' fraction of traffic traveling to the bottleneck.

Moreover, both methods apply state feedback controllers of which the feedback gains need to become properly tuned to guarantee stable system behavior. To this aim, the proposed tuning approach enables systematic evaluation of the system stability behavior and optimal tuning of the gains to minimize the system error (i.e. the difference between a buffers' actual and target state) as fast as possible.

Chapter 7

The Field Operational Test Amsterdam

The field operational test on integrated network management in Amsterdam (Praktijkproef Amsterdam) is a project focused on the design and implementation of an innovative system for the coordinated deployment of traffic management measures in a regional road network. This chapter elaborates on the background of the project, the typical network characteristics and the traffic flow situation. It is discussed how the recurrent and non-recurrent traffic problems at the freeway and urban network lead to different forms of coordination that can improve the network performance. Three types are distinguished, being 'Coordinated ramp metering' within the freeway arterial, 'Coordinated intersection control' within the urban arterials, and 'Integrated control between a ramp meter and its upstream intersections' at the connections. All monitoring and control functions have been integrated in a generic functional architecture and its functionality is illustrated by means of a simulation test. After the build of the production system and a period of intense testing, the system has been operationalized for a period of 11 weeks. The key evaluation results and lessons learned are discussed at the end of the chapter.

This chapter is based on work published in:

- Hoogendoorn, S.P., R.L. Landman, J. van Kooten, M. Schreuder, R. Adams, Design and Implementation of Integrated Network Management Methodology in a Regional Network, *Transportation Research Record*, Vol. 2489, pp. 20-28, 2015
- Hoogendoorn, S.P., R.L. Landman, J. van Kooten, M. Schreuder, Integrated Network Management Amsterdam: Control approach and test results, In *Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems*, pp. 474-479, 2013

7.1 Introduction

Dynamic traffic management measures on a local level, are a widely applied means to increase the road network performance and travel time reliability. It is well known that the uncoordinated deployment of measures can yield suboptimal, perhaps even counterproductive impacts. For recurrent traffic problems as well as for non-recurrent problems (e.g. incidents, roadworks and events) dynamic coordination of traffic management measures is put forward as one of the most promising solution directions for more efficient utilization of road infrastructure.

As we have seen in Chapter 2, many of the network management approaches proposed in literature are based on optimal and model predictive control to maximize the network production or to minimize emissions Schreiter et al. (2012); Zuurbier et al. (2006); Messmer & Papageorgiou (1995); Kotsialos & Papageorgiou (2004); Kotsialos et al. (1999); Zegeye et al. (2009). Despite the fact that optimization based approaches are superior in terms of flexibility and robustness, for practical applications the computational complexity and lack of transparency are sufficiently serious disadvantages to consider heuristic but still generic approaches. However, the few heuristic approaches that are employed in practice such as Wang et al. (2009, 2010), often do not properly target the phenomena that decrease the network production or are very cumbersome to deploy and maintain. For these reasons, there is still limited practical experience on integrated network management, which indicates the need to further explore its potential by means of *field operational tests*. Two paths can be followed:

- Science could try to gather the required insights for solving all foreseen problems, up to the point where operationalization of the system seems trivial;
- A system can be operationalized that is able to prevent traffic problems up to a certain extent, but that can be made more efficient over time by solving encountered research questions along the way.

There is much to say for the latter option, given the fact that our world and future are volatile, uncertain, complex and ambiguous. Furthermore, the design and realization of such systems in practice is a complex task that is

constrained by time and money. This implies that in order to successfully realize an operational system at all, concessions probably need to be made on aspects such as: the system design, the theoretical underpinning of design decisions, full understanding of the system dynamics and the implementation process.

This chapter elaborates on the designed control system within the Field Operational Test Amsterdam. The aim of the project, is to gain experience with integrated network management in practice, to learn about its potential and to better understand the challenges involved based on a learning-by-doing way. The focus in this chapter has been put on the design of the control framework, and in the next chapter the underlying design process is discussed in more detail.

In order to reduce complexity and to make the framework suited for adopting different control or monitoring units, a modular and hierarchical system has been developed. In other words, the system can be easily tailored to different network layouts that have their own typical problems. Moreover, the system needs to be able to autonomously prevent the traffic phenomena that decrease the network performance while network utilization is realized in line with the policy objectives of the road authorities.

7.2 Background of the Field Operational Test Amsterdam

The heavy traffic conditions on the Dutch road network, especially those around the large cities during peak periods, are good reason for our government to invest in solutions that mitigate existing and future mobility problems. In this section we will briefly introduce the policy background of the Field Operational Test Amsterdam and the regional network layout available for the test.

7.2.1 Goals of the government

The goals for traffic and transport in the Netherlands up to 2020 are described in the 'Nota Mobiliteit'. To cope with the current and anticipated traffic problems the Nota Mobiliteit presents the ambitious goals: 'faster, cleaner and safer door to door mobility'. Three interrelated pillars are identified to help realizing these goals knowing: *Building road infrastructure*, *Pricing of infrastructure use*, and *Utilization of infrastructure*. The pillars 'Building' and 'Pricing' are powerful instruments to solve congestion problems, but they have their limitations in the short term. For one, due to air quality regulation and deferred maintenance, building and widening of roads is delayed. Since the implementation of road pricing has been delayed due to technical reasons and the absence of public support, and because congestion is still steadily increasing, the Dutch Ministry formulated a clear policy on the transition towards optimal utilization of the Dutch road network by means of dynamic traffic management of Transport Public Works & Management (2008). This policy is supported by the willingness to structurally invest in dynamic traffic management, given that we can prove the cost-effectiveness of the measures in general, and in particular of the coordination and integration of measures.

7.2.2 Investment in better utilization of road infrastructure

To investigate the potential of integrated network management in reality, the Dutch Ministry of Transportation has provided a budget of 50 million Euros to implement network management in and around the city of Amsterdam. The overall approach is based on smart integration of established concepts, theories and best practices of local and coordinated dynamic traffic control strategies. It was decided to go for stepwise development and implementation of a modular system that can be easily tailored to the prevailing problems within a regional network. If the field operational test is successful, then the aim is to implement the concept in other urban regions as well. Furthermore, contrary to the currently applied scenario-based coordination of traffic management measures, an important prerequisite is that the system will autonomously (i.e. without a man in the loop) define control actions.

The large-scale field operational test will last about four years, including design and development, implementation and evaluation. It considers the freeways, as well as urban roads and arterials. The system will be thoroughly evaluated from a traffic operations and from an organizational perspective. The results are expected to be of high value for national and international road authorities, researchers and industry, showing under which circumstances and to what extent integrated network management is a cost effective solution for congestion.

7.2.3 Situational description

The urban and freeway road network in the Amsterdam region is given in Figure 7.1 and the field test area covers about 175 km^2. The A10 ring road around Amsterdam facilitates large flows of traffic in north, south and eastbound direction. Given their traffic loads and daily traffic problems, the focus has been put on the freeway arterials A10 west and A10 south. In the north of the A10 west, the Coentunnel is located underneath the 'Noordzeekanaal'. The connections between the urban and freeway network along side the A10 west and south are numbered s101 to s113, starting with the s101 just below the Coentunnel at the A10 west. Traffic movements into and out of the city are facilitated by urban arterials that lie perpendicular to the A10 ring road. All connections are equipped with ramp metering installations, with the exception of the s108 where the metering functionality is realized by the intersection signal groups that feed the ramp. Most of the intersections are controlled by vehicle responsive control schemes and over 75 variable message signs are applied to redistribute traffic over available route alternatives (see Figure 7.2).

The most important applications to communicate with the traffic management measures are the CDMS (Central VMS Management System - translated from Dutch) for control of Variable Message Signs, and the CVMS (Central Intersection controller and ramp metering Management System - translated from Dutch) for control of intersections and ramp metering installations.

With respect to traffic monitoring, the speed and flow profiles of freeway traffic are measured by means of loop detectors every 500 meters. Ramp metering installations have their own loop detectors installed upstream and

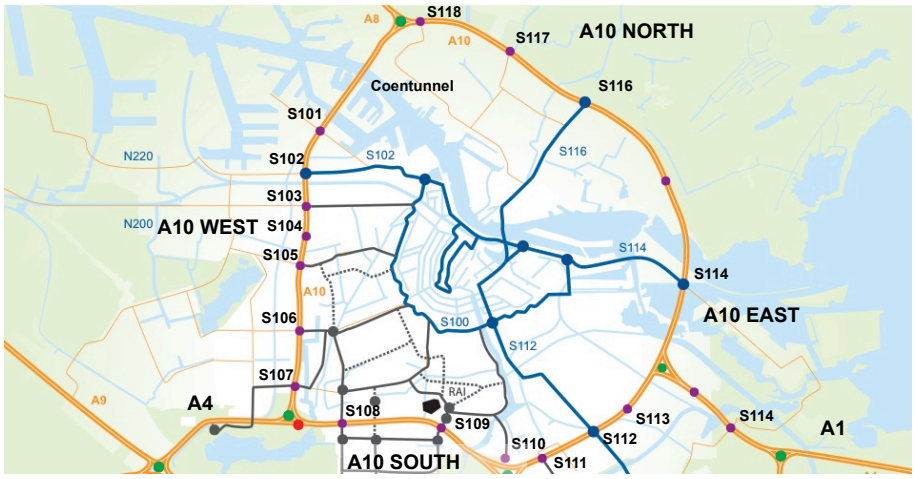

Figure 7.1: Overview of the Amsterdam road network.

downstream the merging area and at the beginning and end of the on-ramp. The different signal directions at the controlled intersections are equipped with loop detectors that measure occupancy and vehicle counts. Detectors in the urban roads can be made available for the estimation of queue lengths.

Two traffic control centers monitor the network and distribute traffic information, one for the freeway network, and one for the urban network. Three authorities are involved in the large field test: the Ministry of Public Works and Water Management, the Directorate-General for Public Works and Water Management and the Region of Amsterdam.

7.2.4 Typical problems in the Amsterdam network

Recurrent problems at the freeway can be found by evaluating the average traffic flow conditions in terms of speed, flow and density over a longer period of time. The recurrent congestion at the A10 ring road is typically caused by a lack of mainstream capacity and by the limited capacity of the off-ramps. For the Field Operational Test Amsterdam we identified the following bottlenecks as can be seen in Figures 7.3a and b:

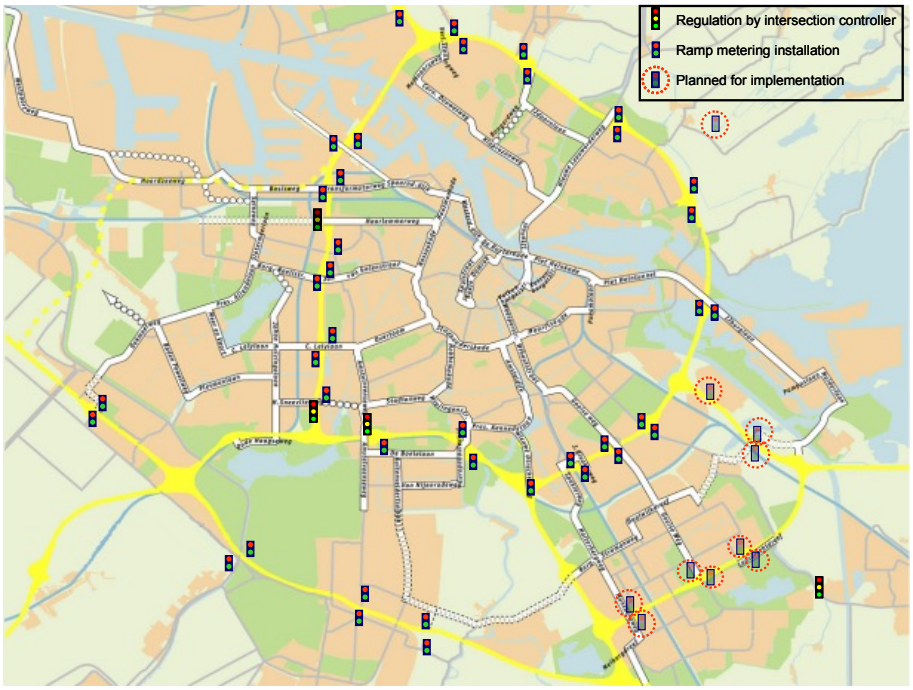

Figure 7.2: Available ramp metering installations in the Amsterdam region.

- In northbound direction the most dominant recurrent bottleneck is the Coentunnel during the evening peak. A similar recurrent disturbance can be seen at the s104 during the morning peak period;
- The recurrent freeway congestion located at the connection s109 in northbound direction is caused by spill-back of the off-ramp queue;
- In southbound direction bottlenecks are visible in the merging areas along the A10 west and at the location where the A4 and A10 west merge into the A10 south;
- Another recurrent congestion is located further downstream at the merge of the s111 during the evening peak period.

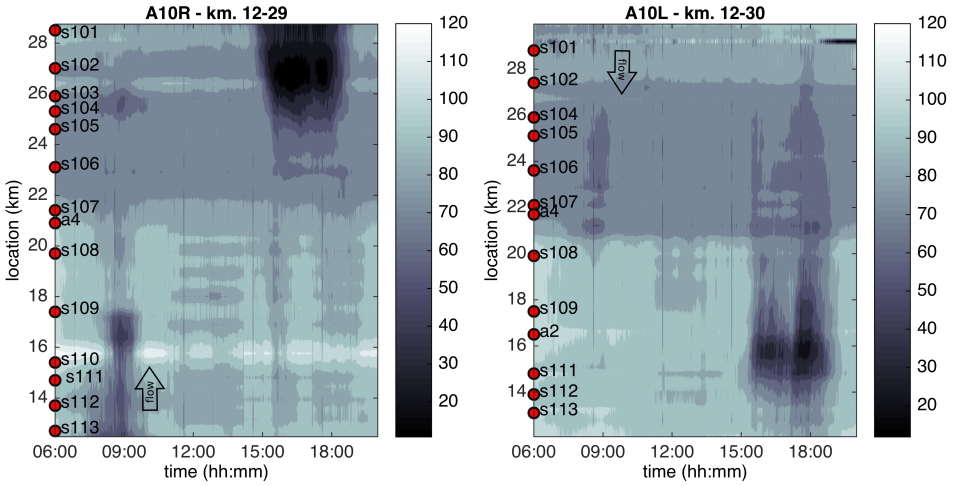

Figure 7.3: Average speeds for the A10 west and A10 south in (a) northbound direction and (b) southbound direction. The plots aggregate the time-space information of November 2011.

These jams cause decreased network outflow due to the capacity drop at the head of the jam and blocking back over upstream bifurcation points. The same holds for the many non-recurrent events that impact the network performance. However, these are not visible within the average flow profiles. Traffic conditions have therefore also been evaluated per day over one month in combination with sources such as incident databases. These evaluations

indicated that incidents happen regularly at the A10, also causing jams that spill back over upstream-located connections.

During the morning peak period large increases in travel time occur within the urban arterials as can be seen in Figure 7.4 for the s116. The queues can spill back to upstream bifurcation points, limiting the network outflow. Similar conditions hold for many of the other urban arterials, especially those connected to the A10 west corridor.

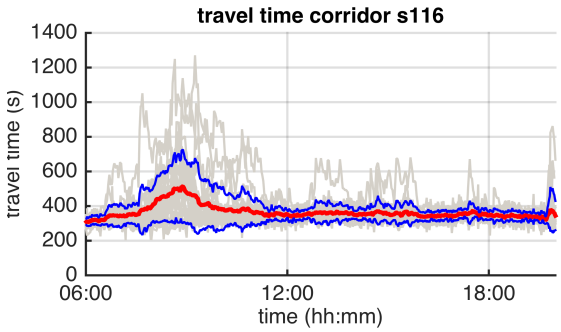

Figure 7.4: Average travel times at urban corridor s116 in direction of city centre.

7.2.5 Test site and typical solution directions

In this section the considerations are shortly discussed that made the A10 west in northbound direction the target area for the field operational test (see Figure 7.5). Moreover, the integrated and coordinated control strategies that have been developed, are based on the recurrent and non-recurrent traffic problems in this area.

Northbound direction

Both the recurrent and non-recurrent bottlenecks (s101, s104 and incidents) at the A10 West in northbound direction cause flow breakdowns with an associated capacity drop and spill-back to upstream bifurcation points. Flow

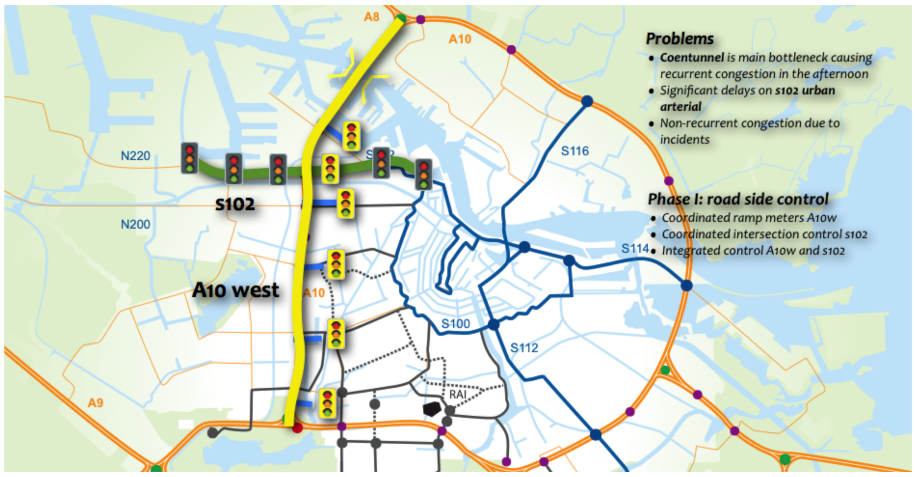

Figure 7.5: The different forms of coordination and integration between control measures derived from the typical bottleneck situations in the Amsterdam network.

breakdown and congestion spill-back can be postponed by temporarily storing traffic upstream of these bottlenecks. *Coordinated ramp metering* can in this respect be an effective solution, since all ramps along the A10 west are equipped with ramp meters, the ramps are located relatively close to each other, and they have a strong downstream orientation.

Traffic can also be stored within the urban arterials upstream of the ramps, which especially holds for the s101 and s102. Both the connections and the further upstream located intersections are controlled and traffic is strongly oriented in northbound direction (i.e. the recurrent bottleneck at the s101). Storing vehicles at upstream intersections can be done by means of *Integrated control between the ramp meter and its upstream intersections*.

Within the arterials travel times strongly increase during the peak periods due to queuing. To prevent queues from spilling back to upstream intersections, *Coordinated intersection control* can be an effective means to stabilize the queues by storing vehicles upstream of the bottleneck.

Southbound direction

Control solutions for the recurrent jams in southbound direction are less trivial. The congestion at the merge of the A4 and A10 west is located relatively far downstream from the ramp metering installation at the s107. Large control delays would make it difficult to properly keep the bottleneck at capacity by means of ramp metering. Moreover, the s107 is a relatively small connection with a weak relation to bottleneck, something that would severely limit the effectiveness with which storage space is used and the bottleneck postponed.

Similar reasoning holds for the recurrent congestion at the s111; there is only the s109 that can be effectively applied to assist in the ramp metering task on the bottleneck. Note that the A2 is a connecting freeway and not a controlled connection. To conclude the s108 does not have a real ramp metering installed, but an intersection controller with ramp metering functionality (see Figure 7.2). Updating the intersection's outflow to the freeway only once during a complete cycle, makes accurate assistance in a coordinated ramp metering scheme impossible.

7.2.6 Control framework and its paradigms

This section explains the conceptual framework, the corresponding components, and the control paradigms of the designed integrated network management system. The system consists of *Monitoring units*, *Control units* and *Supervising units* that are operating on different levels of network scale. The considered network scales are: *Points*, *Arterials* and *(Sub)networks*.

An important control paradigm that has been used in the Field Operational Test Amsterdam is *'solve problems locally if possible, more globally if needed'*. The designed system is hierarchical and gradually escalates to higher levels of control if problems cannot be effectively solved by means of local control. With respect to control units this implies, that when storage space runs out in the direct surrounding of the bottleneck, increasing levels of coordination and integration between control measures are activated to

enable the use of storage space upstream of the bottleneck. The following example illustrates the escalation process for the application of ramp metering:

- As the flows at the freeway are steadily increasing at the beginning of the peak hour, the ramp metering installation will spread out the upcoming platoons from the upstream intersection controller to keep the freeway flowing as smooth as possible;
- As the peak hour advances, a higher flow regime will be detected at the freeway and the ramp metering installation will then start restricting the inflow to the freeway by temporarily holding back traffic at the ramp;
- A first step to further postpone a flow breakdown is the activation of coordinated ramp metering so that upstream-located ramps will start assisting;
- To prevent a ramp from saturating, the intersection controllers located directly upstream start holding back traffic at the signal directions that feed the ramp;
- As storage space is running out, multiple upstream intersections can be coordinated to postpone blocking back of intersection queues.

Also the system's monitoring units gather data on different levels of scale. On the one hand, to identify and diagnose undesired traffic conditions and, on the other hand, to feed the control units with the required input data. A typical example of an useful monitoring unit that is active on a network level is the Network Fundamental Diagram. In Section 7.2.7 a more elaborate overview and description is given of the involved monitoring units.

Archetype situations

With respect to diagnosing undesired network conditions, different archetypical problem situations have been defined that capture all types of bottlenecks that can occur in the target area of the Amsterdam network. These situations are illustrated in Figure 7.6. For each archetype situation specific integrated control approaches have been developed. Let us discuss them briefly:

- **Situation 1:** There is a bottleneck within the urban arterial, which is not hindering traffic on the freeway. The general control idea is to prevent blocking back of queues by holding back vehicles at upstream located intersection arms;

- **Situation 2:** An off-ramp queue spills back to the freeway. The capacity at the controlled downstream end of the ramp is increased by taking away capacity from other intersection arms;

- **Situation 3:** The bottleneck is located on the controlled freeway arterial. The congestion will be postponed or resolved by reducing the inflow to the freeway by means of (local or coordinated) ramp metering, and by integrating the ramp meters with their upstream located intersection controllers;

- **Situation 4:** The bottleneck is located downstream of the controlled freeway arterial, while the congestion spills back onto the controlled freeway arterial. The tail of the queue is stabilized by means of (local or coordinated) ramp metering to prevent further spill-back over upstream off-ramps.

System architecture

To be able to identify the different situations described in the previous section, and to act on these properly, several functions have been developed for state estimation and prediction. This includes functions for bottleneck detection on the urban and freeway network and to act accordingly by deploying the available control units. These functions and their relations are captured in a modular architecture of the integrated network management approach as shown in Figure 7.7. The figure shows four types of functions: *detection*, *monitoring and diagnosis*, *supervision* and *control*. In upcoming sections we will further elaborate on the specific monitoring and control units.

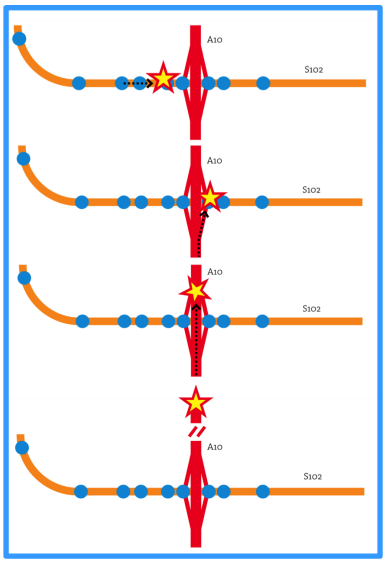

Figure 7.6: The four archetype traffic problems or situations that capture the dominant types of bottlenecks that can occur in the target area of the Amsterdam network.

7.2.7 Monitoring units

Monitoring units continuously monitor the actual traffic situation on the roadway sections in the network. Their function is to prepare all required information for making control decisions on a network level such as: State estimates of the freeway and urban network, anticipated locations and characteristics of bottlenecks, and the available capacity within the network. They retrieve their data from assigned sensors (e.g. induction loops or cameras) and process the data if needed. Hence, they describe monitoring functions rather than physical sensors. This increases the flexibility to remove and add new state estimation techniques without changing the entire system framework. To conclude, the units are defined on a specific clustering level (point, string, subnetwork, network) as can be seen in Table 7.1. Note that some of them provide directly measurable quantities, while others provide indirect quantities.

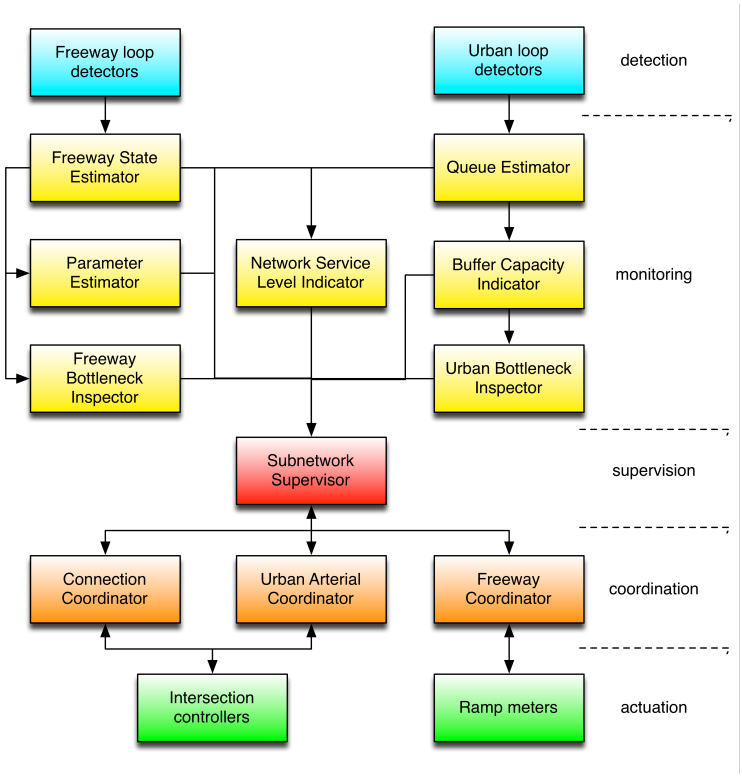

Figure 7.7: System architecture consisting of the designed monitoring, control and supervising units that are used to operationalize integrated network management in practice.

Table 7.1: Examples of different monitoring functions that have been developed for making state estimations of the freeway, the urban and the overall network.

Level	Function of the monitoring units
Point	Macroscopic flow variables speed, flow and density.
	Delays or waiting times
	Bottleneck capacities
	Turn fractions at bifurcations
	Traffic distribution over lanes
String	Queue lengths and storage space estimation
	Travel times
	Remaining route capacity in terms of flow or storage space
	Freeway breakdown probabilities over complete freeway stretch
(Sub)network	Traffic performance in terms of total time spent
	Remaining capacity / storage capacity in subnetwork
	Total inflow and outflow from the network
	Accumulation of vehicles in the network / average vehicle density

Freeway State Estimator

The *Freeway State Estimator* uses freeway loop detectors (placed at approximately every 500 meters on the A10 West) to calculate estimates of the speed, flow and density. To this end, the Freeway State Estimator first corrects for the biased speed observations by applying the CHECK algorithm van Lint et al. (2009). The bias is caused by the fact that arithmetic mean speeds are collected rather than harmonic mean speeds, yielding the average speeds unsuitable for computing densities. After the speeds are corrected, the Adaptive Smoothing Method Treiber et al. (2011) is used to estimate the freeway traffic state at the locations between the loop detectors. Figure 7.8 illustrates its working given the raw data stemming from an incident scenario; the Figures show the correction of missing data by smoothing the observed traffic states over time and space.

Freeway Bottleneck Inspector

Based on the freeway state estimates, the *Freeway Bottleneck Inspector* determines the (predicted) location of a freeway bottleneck. This is achieved by first determining hot-zones (locations where the breakdown probabilities

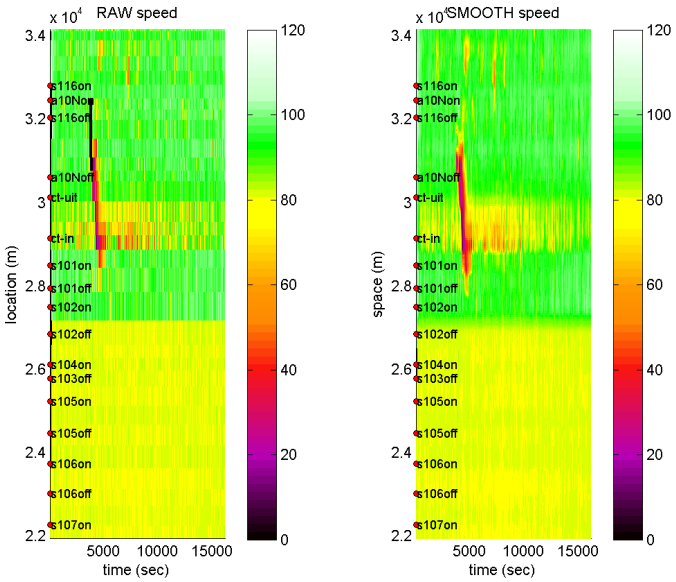

Figure 7.8: Example of the Freeway state estimator input and output, with (a) the raw traffic flow data collected at the loop-detectors, and (b) the output in the estimated freeway traffic state based on correct smoothing of data over time and space. Data shows the situation in case of an incident simulated in Vissim.

are high) using historic data, and combining this with prevailing speed estimates. Using this method allows the prediction of breakdown three minutes ahead of its actual occurrence with sufficient reliability. This does require extensive training of the procedure and only applies to recurrent bottlenecks. For incident situations, historic data provides little information about the bottleneck probabilities. In this case, the Freeway bottleneck inspector is reactive, showing the current characteristics of the queue (see Figure 7.9). The location and characteristics of an identified bottlenecks are shared with the governing supervisor and control units.

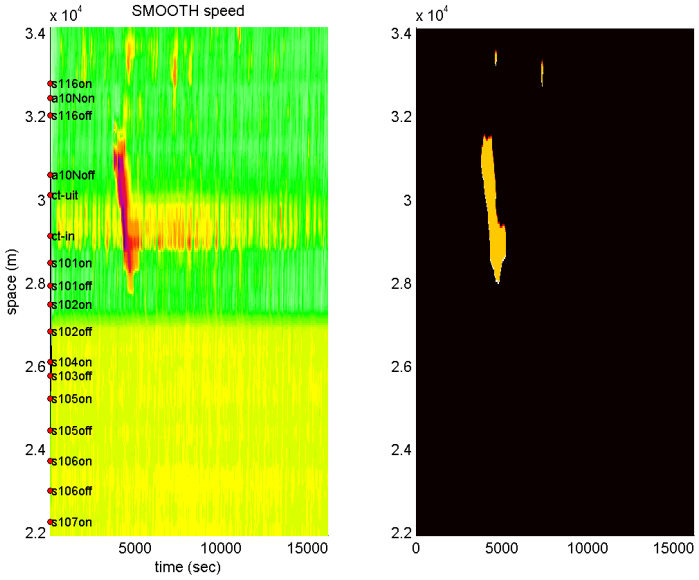

Figure 7.9: Example of the Freeway bottleneck inspector input and output, with (a) the smoothed Freeway state estimator results serving as input and (b) the outcomes the Freeway bottleneck inspector in terms of the head, body and tail locations of congestion (simulated in Vissim).

Parameter Estimator

The *Parameter Estimator* estimates the critical density (i.e. the capacity) of all freeway segments and bottleneck locations based on real time traffic flow measurements. It essentially searches for the top of the fundamental diagram

of each freeway segment, accounting for conditions that influence it such as the weather, infrastructure characteristics and traffic flow composition. Up to date estimates of capacity enable full utilization of the freeway capacity, i.e. to control it such that overloading and underutilization of the freeway infrastructure can be prevented by means of control. To this end, the Kalman filtering approach described in Smaragdis et al. (2004) is used.

Queue Estimator

For the urban arterials, different monitoring functions have been developed. The *Queue Estimator* estimates and predicts the queues at the off-ramps, on-ramps and the intersections using a variety of estimation and prediction techniques. The approach selects the best queue estimation for a particular situation, and provides it to the Urban Bottleneck Inspector and the Buffer Capacity Indicator.

Urban Bottleneck Inspector

Based on the queue estimates, the *Urban Bottleneck Inspector* determines (or predicts) if there is a bottleneck at the urban arterial. An urban bottleneck is defined by a queue causing spill-back problems or by an over-saturated direction at an intersection. The queue estimates are also used to determine the remaining buffer capacity.

Buffer capacity indicator

To this end, the *Buffer Capacity Indicator* compares the current queue lengths to the allowed buffers and returns the difference. Note that the buffers, describing the locations and lengths of the queues that are allowed, given the prevailing network level-of-service, have been determined in close collaboration with the municipality of Amsterdam.

Network service level indicator

The *Network Service Level Indicator* computes the average level-of-service of a network. This is achieved using the generalized Network Fundamental Diagram (g-MFD)[1]. It relates the average network density and the spatial density variability to the network production. An example of the g-MFD is given in Figure 7.10, showing how the different levels-of-service (green, yellow, red, and black) could be defined as a function of the accumulation and its spacial variance. The graph also enables the identification of remaining storage space in the network before its performance will decrease.

7.2.8 Supervisors

Supervisors have central knowledge on the physical network layout and the demand characteristics. Hence, they are able to interpret the actual network state by combining this knowledge with the real-time information coming from the monitoring units. Let us briefly discuss some of the supervisor's functionalities.

- **Interpretation.** The primary task of a supervisor is the interpretation of the network states and archetype problems based on traffic data coming from the monitoring units;
- **Changing objectives.** When the network degrades in overall performance or when specific traffic problems arise, the supervisor can decide to change the control objectives and functions. For instance, based on the archetype traffic situation and the current network level-of-service within the network, buffer configurations and buffer sizes can be adjusted;

[1]Recall from Daganzo & Geroliminis (2008); Geroliminis & Daganzo (2008) that the MFD describes a crisp relation between the accumulation and the overall production of a corridor or (sub)network. The production of the considered network area is derived by weighting the measured flows over the lengths of their roadway stretches. If the network accumulation is less than the critical accumulation, no serious problems in terms of traffic operations have occurred. When the accumulation is, however, larger than the critical accumulation, the average network performance will start to decrease and a problem will be identified.

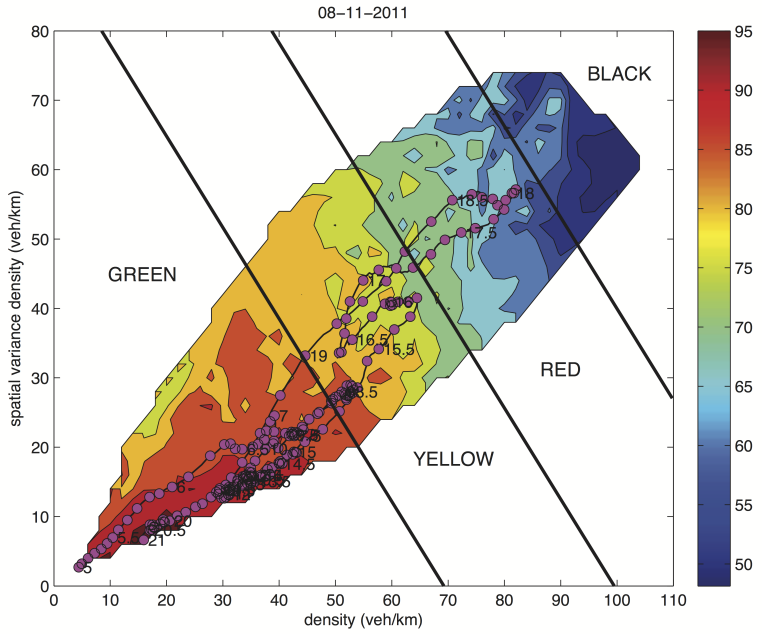

Figure 7.10: The network level-of-service based on the average density and the spatial variance in the density. The figure shows the average speed on the A10 and the (related) areas where the different levels-of-service hold. The figure also shows the path of the density and its standard deviation for a particular day, showing how the network level-of-service evolved.

- **Data processing.** Its overall knowledge about the network layout and typical demand characteristics enables the preparation of important control inputs. The effective space available for control purposes can be determined by accounting buffers' total available space with the turn fractions to the active bottleneck;

- **Delivering control inputs.** Based on the active situation(s) and conditions within the network, the supervisor supplies the control units with inputs such as: bottleneck locations, bottleneck characteristics, control parameters and up-to-date freeway capacity estimates.

One such supervisor has been operationalized during the field operational test, being the *Subnetwork Supervisor* governing the freeway arterial A10 west in northbound direction, the complete urban arterial s102 and the connections s104 to s107. Determining the relevant buffer configuration is an important task of the subnetwork supervisor. The buffers describe how much traffic may be stored within the buffers. This is dependent on the level-of-service and the traffic situation at hand, as Table 7.2 below indicates. For instance, in case of situation 3 (bottleneck on the A10), arm 18 can have a queue of 40, 60, 100 or 100 meter, dependent on the prevailing level of service (green, yellow, red or black respectively).

Figure 7.11 illustrates the maximum storage space at the intersection arms upstream of an on-ramp given the active service level. It can be seen that not all legs of the intersection are used to the same extent for all situations. This has to do with the effectiveness of using the storage space (i.e. which percentage of traffic is going to the on-ramp), or it may have a policy background (is the arm on an important urban arterial that requires high throughput).

7.2.9 Control units (local and coordinated)

The control units influence the traffic flows by means of physical actuators (ramp meter installations, intersection controllers, variable message signs, etc.). They are defined at different levels of scale, being: *point*, *string* and *(sub)network*. There are 4 Input/Output (I/O) relations defined for the control units:

- Inputs coming from the monitoring units;

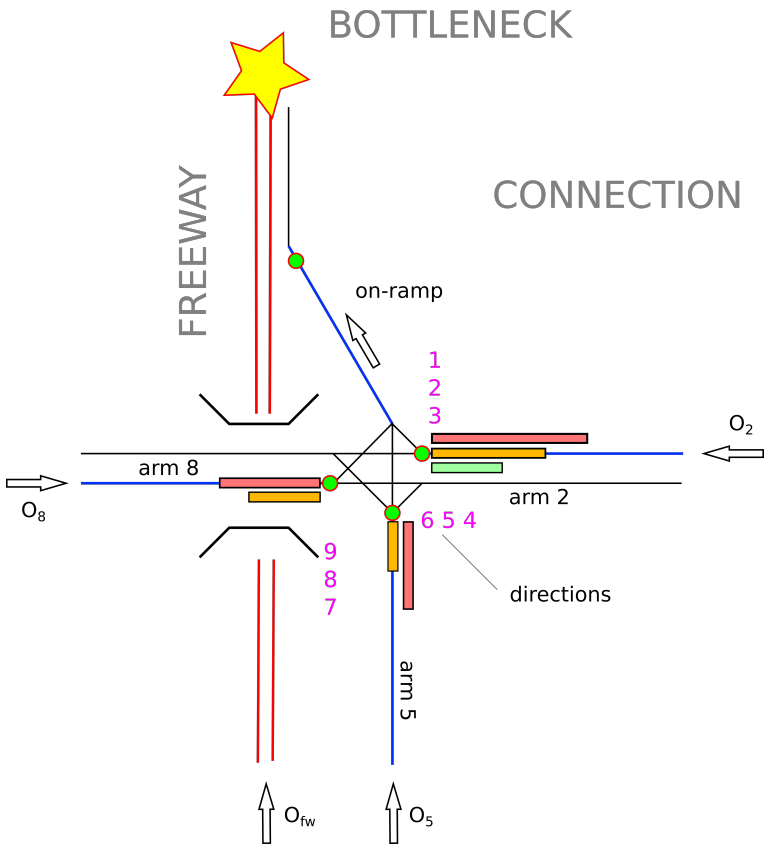

Figure 7.11: Traffic can be stored on the on-ramp and on the controlled intersection during this archetype situation 3. The size of the intersection buffers is dependent on the active level-of-service and indicated by means of colors.

Table 7.2: Buffer lengths (in meters) per storage space (17, 18, 19) for different combinations of levels-of-service (1,2,3,4) and archetype traffic situations (1,2,3,4).

Buffer		Situation 1	Situation 2	Situation 3	Situation 4
17	LoS 1	25	100	40	60
	LoS 2	50	100	60	80
	LoS 3	-	100	100	100
	LoS 4	-	100	100	100
18	LoS 1	25	100	40	60
	LoS 2	50	100	60	80
	LoS 3	-	100	100	100
	LoS 4	-	100	100	100
19	LoS 1	25	50	-	-
	LoS 2	50	80	-	-
	LoS 3	-	100	100	100
	LoS 4	-	100	100	100

- Inputs coming from higher level control units or supervisors;
- Outputs to the physical actuators;
- Outputs to higher level control units.

The control units at the lowest clustering level (points) perform autonomous when problems in the network occur at a local level. When the problems become more severe, the higher order control units (i.e. coordination algorithms) take over and the level of coordination increases. To this aim, local control units are equipped with the proper level of intelligence to communicate with higher order control units that realize coordination, and with supervisor units that prepare control inputs.

Coordination algorithms divide the control tasks over the coordinated and integrated lower level control units, based on their ability to control the situation at hand (available space or potential effect) and their distance to the unit in need of assistance. The latter determines how long it takes before the assistance has an effect. The complexity of the search space is reduced by the predefinition of fixed combinations of coordinated and integrated control measures. The lower level units then no longer function autonomously, because they are receiving instructions while constantly being monitored and

evaluated by a higher level unit. Hence, units comply with the instructions as long as these do not cause traffic problems at a local level. For instance, if a coordinated ramp metering scheme would instruct a ramp to increase its outflow, while that would result in a flow breakdown, then the local control signal will be realized. The list in Table 7.3 provides a number of examples of the different control functions that have been considered within the project.

Many of the above mentioned functions can be realized by means of local dynamic traffic management measures. For instance, freeway flow breakdowns can, up to a certain extent, be prevented by means of ramp metering and the throughput (use of capacity) at a local intersection can be optimized by means of intersection control. However, functions such as preventing ramp saturation, preventing spill-back in urban corridors or distributing metering tasks over multiple ramps need integrated and coordinated control algorithms. Such control units will be called Coordinators in the remainder of this chapter. The *Freeway Coordinator* and the *Urban Arterial Coordinator* that have been designed and operationalized in the field operational test will be discussed in the remainder of this section.

Freeway coordinator

The *Freeway coordinator* coordinates the ramp meters along the freeway arterial, in order to use upstream ramps to longer postpone a freeway flow breakdown. Before looking at the specifics of the coordination algorithm, let us first have a look at some of the objectives this higher level control unit pursues in combination with the lower level units (ramp meters) it coordinates. Typical functions of local and coordinated ramp metering are:

- Reducing the inflow to the freeway at one or more ramps to postpone the onset of congestion or to resolve it;
- Facilitating the merging process by smoothing the ramp flow to the freeway (platoon dispersion);
- Making the use of the freeway less attractive by increasing the waiting times at the controlled ramps.

Table 7.3: Examples of different control unit functions.

Level	Function of the control unit
Point	Facilitating weaving process at merges and weaving areas. Preventing freeway flow breakdown by reducing the flow from urban network to freeway. Preventing ramp saturation by reducing the flow from the urban network to the ramp. Creating extra capacity for merging flows by means of dynamic lane closure. Maximizing throughput at intersections.
String	Postponing freeway flow breakdown by distributing the metering task over multiple ramps. Improving the throughput of major urban corridors by preventing spill-back of queues. Improving the distribution of traffic over lanes by means of dynamic lane allocation. Facilitating weaving process by dynamic speed limits. Controlling the service levels within (parts of) routes in line with policy objectives.
(Sub)network	Improving throughput A10 west network by holding back traffic at the perimeters of the considered network or rerouting over network elements with redundant capacity in terms of flow and storage space.

Below a brief overview is given of the proposed coordination strategy of the Freeway coordinator and how it is interacting with the other monitoring and control units in the framework. In Chapter 5 an elaborate and more formal description of the algorithm is given.

- For individual metering purposes on merges and other bottlenecks the local ramp metering algorithm ALINEA is used based on the critical and current density at the location of the (potential) bottleneck. Ramp metering installations gather their own control inputs (i.e. speed, flow, density and queue lengths) by means of freeway and ramp detectors;
- Ramp metering ability has been extended to all other parts of the freeway. To this aim, the Freeway Bottleneck Inspector identifies the location of the (potential) bottlenecks and the Parameter Estimator defines the freeway capacities that serve as setpoints for the metering algorithm (in terms of critical density). The Freeway coordinator subsequently instructs the ramp metering installation directly upstream the bottleneck to start metering;
- With respect to the realization of coordinated ramp metering, the Freeway coordinator receives from the Subnetwork supervisor up to date freeway and ramp state estimates and bottleneck characteristics. From the coordinated ramp meters it receives the status, local metering rate and parameter settings. The Master status is given to the most downstream located ramp that is running out of space. Assisting ramps are given the Slave status and a coordination metering rate that enables the full utilization of ramp space.

The basic functionality of the coordination algorithm is as follows: First the *saturation duration*[2] of the master ramp is determined. By synchronizing the saturation duration of slave ramps with that of the master, the outflow is restricted when the master ramp runs out of space. This creates gaps in the mainstream flow that travel from the slaves to the master. At the moment that the master ramp detects these gaps, more vehicles can be released from the ramp without causing a flow breakdown. To conclude, saturation durations for the slaves are corrected for the *travel time it takes a gap to travel to the master ramp*, so that the ramps saturate in downstream order and all storage space is effectively used.

[2]The saturation duration describes how long a master ramp meter can still hold back traffic given its metering rate and the storage space left on the ramp.

Urban arterial coordinator and Connection coordinator

The *Urban arterial coordinator* is responsible for the coordination of the intersection controllers to prevent queue spill-back within the urban arterial. The Connection coordinator is a simplified version of the Urban arterial coordinator with the exact same technical functionality, but it only considers the intersections that are feeding the ramp to prevent ramp queue spill-back. For the current implementation these coordinators only deal with situations 3 and 4 as described in Section 7.2.6. This effectively means that ramp saturation is prevented to postpone a freeway flow breakdown. In Chapter 6 a more elaborate and formal description is given of these coordinated and integrated control approaches. Within our framework, the process is as follows:

- The Subnetwork supervisor determines the buffers that can be used and their size based on the archetype situation (i.e. situation 3 or 4) and the network's level of service;
- To fully use the available buffers, the Urban arterial coordinator (or Connection coordinator) synchronizes the relative space of the buffers with that of the ramp;
- When the ramp is saturating this mechanism ensures that the buffers (feeding the ramp) are instructed to lower their outflow;
- The corresponding green time adjustments are communicated to the intersection controllers before the start of the next cycle.

Despite that the functionality has not been explicitly formulated for situations 1 and 2, we can use similar control principles. The Urban arterial coordinator receives in these cases an instruction from the Subnetwork supervisor to stop the queue spill-back from an urban bottleneck or to limit the queue length at the off-ramp. This can be done by maintaining a target value in terms of relative storage space at the critical buffer or off-ramp by redistributing the capacity over the different intersection directions. However, when increasing the flow into the urban arterial, additional care needs to be taken to prevent spill-back at more downstream located intersections.

7.2.10 Simulation example

In this section the potential en functioning of the integrated network management system is briefly illustrated based on a simulation test case. The complete control architecture consisting of all monitoring and control units was prototyped within the Vissim 5.30 simulation environment. It is beyond the scope of this chapter to show all the test, validation and evaluation results.

One of the important insights gained from the simulations, relates to the configuration of the system to control the recurrent freeway congestion at the s101. As discussed before, there are six controlled on-ramps (i.e. s101, s102, s104, s105, s106 and s107) located along the A10 west freeway in northbound direction. The most recurrent congestion generally occurs just downstream of the s101 on-ramp, which in our coordinated ramp metering concept then should become the master ramp meter. All upstream located ramps are subsequently assigned the task to assist in keeping the bottleneck at capacity. As explained in Chapter 5 and 6, the assistance mechanisms for postponing saturation of the master ramp are:

- **To use upstream ramp storage spaces** by synchronizing the saturation duration of the assisting ramps with that of the assigned master ramp;
- **To use upstream urban storage spaces** by synchronizing the relative storage space of assisting urban buffers with that of the assigned master ramp.

To realize adequate assistance by means of the coordination algorithms, the conditions at the master should not fluctuate too heavily, because then the state and its dynamics are difficult to track. This typical problem manifests in the Amsterdam case; the buffer space on the s101 is very small due to the short on-ramp that only allows space for a few vehicles. This implies that it fills up and empties very quickly due to changes in the ramp metering rate. These highly fluctuating conditions make it impossible for the coordination algorithms to properly assist in the metering at the moment the s101 is coined master.

To circumvent this problem, the coordinated ramp metering algorithm was configured such that the s102 (the ramp upstream of the s101 with more storage space) was assigned as master. The s101 would retain its local me-

tering functionality. It turned out that in this case the coordination scheme becomes effective. As can be seen in Figure 7.12, the controlled case results in lower congestion densities, much less spill-back and a shorter total congestion duration. Especially the latter implies that the impact of the capacity drop is strongly reduced. Moreover, limiting the spill-back to upstream located off-ramps also has a positive effect on the network outflow.

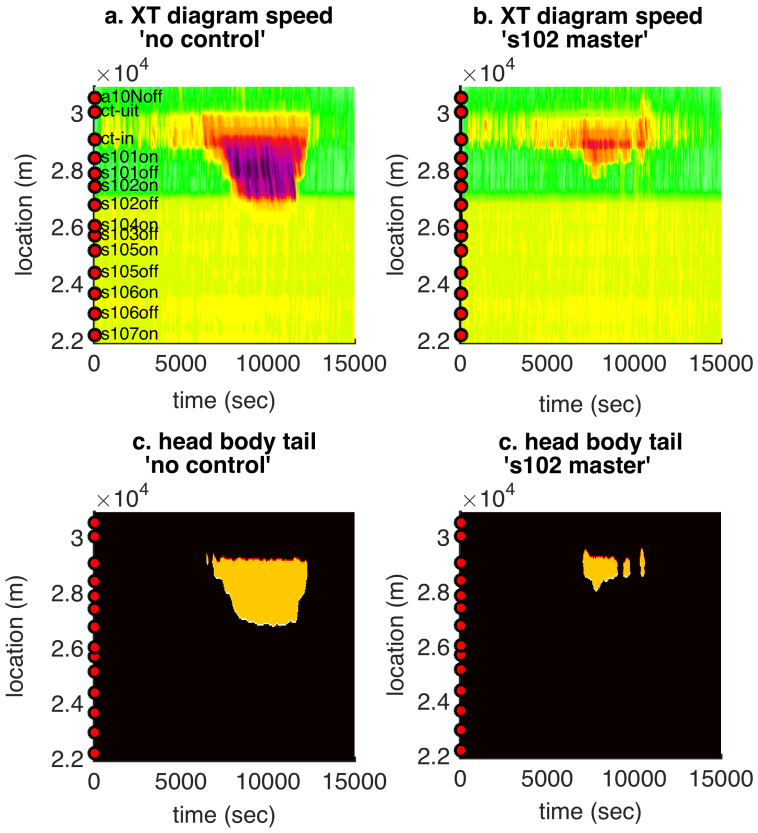

Figure 7.12: Overview test case results with a, b) the time-space diagrams indicating the freeway speeds for the no-control case and control case respectively, c,d) the head, body and tail indications of congestion for the no-control and control case respectively.

In Table 7.4 an overview is given of the start and end time of the congestion, its duration and the maximum queue length. This is the outcome of a single simulation run; due to the stochastic nature of the simulation, outcomes will vary from the one run to the next.

Table 7.4: The congestion characteristics during the no-control and control case of the simulation test case.

Scenario	start congestion (s)	end congestion (s)	duration (min)	length congestion (m)
no control	6450	12270	97	2446 m
s102 master	7080	9000	41	1198 m

To conclude, in Figure 7.13 it is shown that all coordinated ramps are holding back traffic to assist in the metering task on the freeway bottleneck. There can be various reasons why the buffers at upstream located ramps do no become fully utilized. For instance, ramps with small demand will have difficulty in synchronizing their state with that of the master; the maximum realizable outflow reduction of an assisting ramp might not be sufficient to fill at the same rate as the master does. This also causes assisting ramps to quickly resolve their ramp queue at the moment they receive the incentive to release traffic.

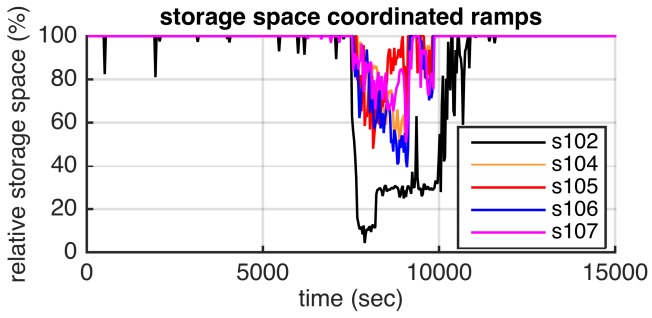

Figure 7.13: Storage space utilization of the coordinated ramps during the control case of the simulation test case.

7.3 Results of the field operational test

The system has been operational and its performance evaluated. Some of the key findings will be discussed in this section and related to the work in this thesis. The pilot took place from the 14th of April 2014 to the 27th of June over an 11 week period in which the system was active every even-numbered week and inactive every odd-numbered week. When the system was inactive, the normal traffic management system was operational (i.e., ramp meters with local metering functionality and a fixed-time coordinated network controller on the s102).

A distinction is made between delays on the urban network and ramps and the freeway. The complete overview of the results can be found in Beenker et al. (2015). Although the developed approach is generic with respect to the different kinds of bottlenecks that can occur, during this testing phase the key objective was to delay freeway congestion and reduce the recurrent congestion effects caused by the s101 and the Coen Tunnel.

7.3.1 Effects on Freeway A10

Figure 7.14 shows the delays on the freeway per minute over the day. Note that the collective delays over the entire peak are identified by the total area under the graphs of the active and inactive system. The figure convincingly shows the effect of coordination on the freeway flow conditions: the collective delays on the freeway are reduced by an average of 190.3 vehicle hours per peak period.

Looking deeper into the causes for the improvements on the freeway operations, one can conclude that the coordination scheme at the core of the developed method resulted in delaying the onset of congestion by 21 min compared with the situation in which the system was inactive Hoogendoorn et al. (2016). This number was established by looking at the breakdown instants of (on average) 16:27 and 16:06h when the system was active and inactive respectively.

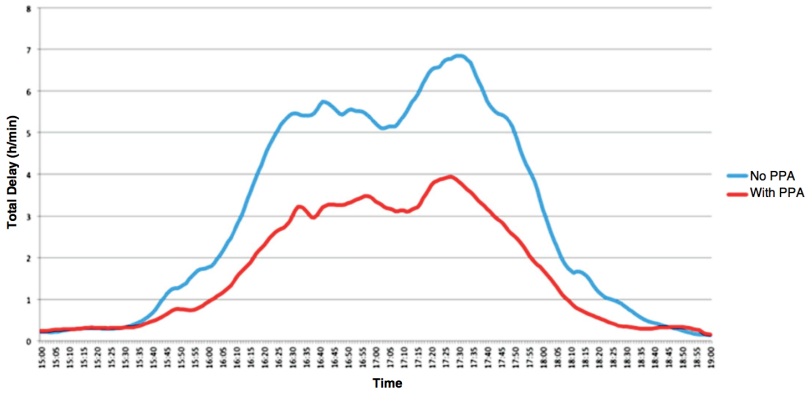

Figure 7.14: Comparison of generated delays during the field operational test in practice for (a) the situation with an active network management system in red, and (b) the situation with the normal traffic management measures active in blue.

7.3.2 Effects on urban network

The system has had a negative effect on the total delays on the urban network and the on-ramps. The total delays increased by 203.8 vehicle hours; 77.5 of the 203.8 vehicle hours were experienced on the on-ramps Beenker et al. (2015). This finding implies that the overall effect of the system was not positive, but neutral. However, a more detailed analysis presented in Hoogendoorn et al. (2016) shows some of the key reasons for the effect not being as expected. Improved tuning and configuration of the system's components are very likely to yield larger improvements, i.e. the impact on the urban network can be reduced, while the positive effect on the freeway is maintained.

The table shows that on average 203.8 vehicles per hour were queued in the buffers, about 105.2 vehicles have been effectively queued. In other words, effectively causing 105.2 vehicle hours delay at the buffers, yields 190.3 vehicle hours saved at the freeway. The concern is then to minimize the delays caused to traffic that is not moving to the bottleneck. An approximate 1:2 ratio is the result, implying that delaying one vehicle for 1 min on the urban network or ramps yields a reduction of 2 vehicle minutes of collective delays on the freeway. A direct consequence of this simple analysis

(A)		(B)	(C)	(D)	(E)
Buffer ID		Fraction	Avg. (Extra) Veh. in Queue		In Config.?
			Actual	*Effective*	
s101	14	100.0%	10.3	10.3	1
s102	11	100.0%	8.7	8.7	1
s102	16	100.0%	1.7	1.7	1
s102	31	92.8%	20.9	19.4	1
s101	24	90.4%	12.7	11.5	1
s101	27	85.5%	2.1	1.8	1
s101	25	84.3%	3.5	3.0	1
s101	28	78.3%	5.6	4.4	1
s104	40	63.9%	9.3	5.9	1
s107	70	62.7%	9.4	5.9	1
s106	*60*	*56.6%*	*17.2*	*9.7*	1
s102	4	51.8%	0.0	0.0	0
s102	8	45.8%	3.6	1.6	0
s102	6	43.4%	5.1	2.2	0
s102	10	39.8%	13.1	5.2	0
s105	50	39.8%	10.4	4.1	0
s104	44	30.1%	0.7	0.2	0
s105	58	30.1%	0.0	0.0	0
s102	18	27.7%	0.1	0.0	0
s104	41	25.3%	0.4	0.1	0
s104	43	22.9%	5.2	1.2	0
s102	3	22.3%	0.0	0.0	0
s105	54	20.5%	0.4	0.1	0
s106	61	18.1%	8.9	1.6	0
s102	17	15.7%	0.1	0.0	0
s105	55	15.7%	1.3	0.2	0
s106	65	15.7%	14.5	2.3	0
s105	56	13.3%	0.0	0.0	0
s105	57	12.0%	13.9	1.7	0
s102	19	9.6%	21.3	2.1	0
s107	71	9.0%	0.9	0.1	0
s102	9	8.4%	2.5	0.2	0
Total (reference):			203.8	105.2	
Total (new config):			101.4	0.0	

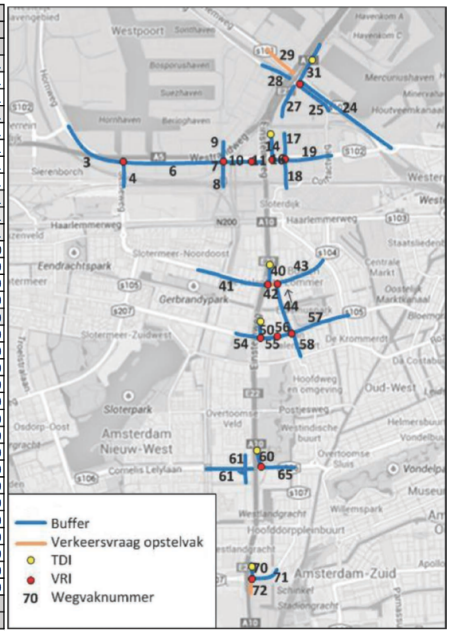

Figure 7.15: An overview of all buffers that have been available for coordination with in columns (A) their ID, (B) their turn fraction of traffic to the bottleneck, (C) the average number of vehicles in the queue, (D) the effective number of vehicles in the queue, (E) indication of more efficient buffer configuration (see Hoogendoorn et al. (2016)).

is also that buffers with a fraction less than 50% are not likely to be effective under the current functioning of the system.

Assuming that the benefits on the freeway scale linearly with this number, the resulting benefits on the freeway can be approximated. When including buffers with a fraction larger than 55%, then the estimated total number of vehicles queued is 101.4 and the number of effectively queued vehicles 88.2. Hence, working with this new configuration is expected to result in $82.2/105.2 * 190.3 = 148.3$ vehicle hours delay at the freeway against 101.4 vehicle hours delay at the urban network.

7.4 Discussion on design aspects

7.4.1 System activation

Anticipating a bottleneck would imply the need of a 3-5 min-ahead prediction of traffic breakdown. In Hoogendoorn et al. (2016) it is stated that the system activated 34 minutes sooner than congestion would occur in the situation without the system. As a result, unnecessary delays on the urban network and the ramps of about 157 vehicle hours in total were caused. This premature activation is assumed to cause buffer space to be depleted too soon. This would have an impact on the time congestion can be postponed, since a fair share of the buffer space has already been used before ramp metering becomes useful.

As we have seen in chapters 5 and 6, both the coordinated ramp metering and the coordinated intersection controllers are of the feedback type. Premature activation does not necessarily mean that undesired delays at the buffers are caused and their buffer space is depleted too soon. There are two direct causes for unnecessarily holding back vehicles, directly after activation, when using feedback type of control:

- **Suboptimal tuning of parameters.** When the initial metering target values at activation are set to low, it will take time for the feedback controller to gradually find a metering value that keeps the bottleneck at capacity. Every control interval in which the resulting metering is too stringent, traffic at the ramps is unnecessarily delayed. Moreover, when the feedback gains are not properly tuned and the gains set too

low, then it will take the controller too long to find a good metering rate. In relation to the first point this implies that even more traffic becomes unnecessarily delayed.

- **Conservative settings of target metering values.** If the setpoints (e.g. in terms of critical density) are defined to conservative, less traffic will be allowed onto the freeway than its true capacity. This results in the holding back of too much traffic and a suboptimal freeway flow. As stated in Hoogendoorn et al. (2016), the target values for ramp metering indeed turned out to be to conservative and for the next phase of the field operational test changes were made in the Kalman filter that is used to dynamically estimate the critical densities. Higher values were found, which lead to a less strict metering strategy and a potential 42% of additional delay saving on the on-ramps and urban arterials.

7.4.2 Buffer protection strategy

A queue protection approach (or, rather, buffer protection approach) was implemented that ensures that when buffers become too full, queues are flushed Hoogendoorn et al. (2016). This means that ramp metering rates are set at their maximum value, and that the intersection controllers give maximum green to the directions in which traffic was buffered. The result is an immediate breakdown on the freeway and the severity of the resulting bottleneck was found to be somewhat higher than when the system was inactive.

A more advanced buffer protection system may yield reduced freeway and urban delays. For instance, buffers can also be kept filled once full; effectively meaning that the incoming traffic demand from the buffer is immediately released on its downstream side. As discussed in chapters 4 and 6, the aim should be to ensure that the buffers are emptied at the moment the freeway congestion is resolved. This will limit the hindrance to vehicles that do not travel toward the bottleneck. The number of buffers and their maximum size should for this reason be carefully chosen.

7.4.3 State estimates for the complete freeway stretch

In order to deal with recurrent and non-recurrent problems at the freeway, a complete state estimate of the considered stretch is required. Such state estimate can be made by implementing many loop detectors in the freeway surface. However, this is a costly alternative. Another way is to estimate traffic states between measurement locations by means of techniques such as the Advanced Smoothing Method Treiber et al. (2011). This method does produce realistic flow, speed and density patterns, but smoothening the truly measured values from loop detectors over time and space, might result in an underestimation of the freeway densities. When applying ramp metering based on these smoothed values, too much traffic might be allowed to enter the freeway, resulting in congestion.

7.4.4 Queue length estimations

The volatile character of queue dynamics and the limited number of loop detectors available to properly estimate queue lengths, make it difficult to identify the true available storage space in the system. To adequately control the stabilization of queues in the network and to fully utilize available buffers, it is recommendable to further investigate cost efficient and more accurate queue estimation techniques (see also in Hoogendoorn et al. (2016)).

7.4.5 Service level indicator

It sounds reasonable to allocate buffers and determine their size based on the service level of the overall network. For instance, when the network is congested, more storage space is made available to solve traffic problems. However, as we have seen in Chapter 4, the allocation of buffers within a coordinated control scheme is situation dependent. Factors like the oversaturated peak period duration and the size of the capacity drop determine the potential benefits of applying coordination, and hence, if it is beneficial to include a buffer within the coordination (i.e. with a certain traffic fraction to the bottleneck). In other words, without thoroughly considering the situation conditions, allocating storage space based on the average network production, may lead to suboptimal control actions.

7.5 Conclusions

In this chapter we have introduced the large-scale field operational test on integrated traffic management in Amsterdam. The designed control framework, its paradigms and its monitoring and control units have been elaborately discussed. The aim of the project, was to gain experience with integrated network management in practice, to learn about its potential and to better understand the challenges involved, based on a learning-by-doing way.

The control framework is based upon the paradigm *'Solve problems locally if possible, more globally if needed'*. If traffic problems cannot be solved by local measures, then higher levels of control are activated. To this aim, different forms of coordination are defined on the level of *strings* and *(sub)networks*, such as *Coordinated ramp metering*, *Coordinated intersection control* and *Integrated control of ramp metering and upstream intersections*. The framework has a modular setup, enabling stepwise deployment of the system, and easy adoption of new sensors and monitoring and control units.

The system aims at preventing the following undesired traffic phenomena: *freeway flow breakdown with associated capacity drop*, *blocking back of queues*, and *making optimal use of available network capacity and storage space*. The workings of the control system and its components has been discussed by means of examples that illustrate the outputs of the system components, a simulation test case in which the complete prototype code of the framework is operational, and the evaluation results of the production system in practice. The results show that the control framework and its individual components can improve performance if setup and configured well.

Chapter 8

Design process for integrated network management

Designing an integrated and large-scale network management system is a complex task. This chapter presents the measures that have led to a gradual and successful implementation of such large-scale and integral control framework within the Field Operational Test Amsterdam. It is discussed how undesired traffic conditions can be diagnosed and how phenomena responsible for suboptimal network outflow can be identified. Moreover, different development phases are defined and elaborated on. As control and monitoring concepts need to be designed based on a thorough understanding of the problems at hand and a clear understanding of potential solution-paths.

This chapter is based on the work performed and experiences gained during the Field Operational Test Amsterdam.

8.1 Introduction

When faced with the challenge to improve the road network performance by means of traffic control measures, there are several reasons to start the process with a thorough traffic analysis. This may sound trivial, but as we have discussed in Chapter 2, it happens quite regularly that control designs are developed without a clear understanding of the phenomena that the system targets and if this is done effectively. Understanding the root causes of the undesired traffic situations:

- Will directly point us towards suitable control solutions;
- Enables tracing back conceptual, technical and functional errors, including sub-optimally tuned parameters in the system design;
- Is a basis for understanding the potential effectiveness of control solutions, which is also important for creating support among stakeholders;
- Is key to underpin the test side location choice and to make relevant investment decisions with respect to infrastructure preparations, such as: road layout, dynamic traffic management arsenal, and communication means between measures and between measures and the road user.

In this chapter, the key steps are discussed that led to successful and stepwise operationalization of the control framework for the Field Operational Test Amsterdam. The proposed process is of the feedback type, meaning that technical or functional errors may lead to conceptual changes. Examples from the design process within the field operational test are given for illustration purposes. Section 8.2 elaborates on how to identify undesired network states. In Sections 8.3 and 8.4 it is discussed how the root causes of the undesired traffic phenomena lead to network-wide control solutions. The process of designing and operationalizing the system is given in Section 8.5, based on the use of models for testing concepts and validating their technical and functional design. The chapter concludes with the conclusions and recommendations in Section 8.6.

8.2 Identification of undesired network conditions

There are different types of undesired conditions within the road network that can be identified as problematic. Most problems are caused by a mismatch between available supply and demand, which is easily recognized by the onset of congestion at the freeway or waiting queues at the urban network. Two types of problems are distinguished:

- **Recurrent problems** are probably the most important to address, given the simple fact that these problems occur on a regular basis and thus solving them will significantly improve the network performance. They become easily visible when looking at traffic data that is aggregated over longer periods of time.
- **Non-recurrent problems** such as incidents or peak demands due to events can also result in severe congestion and a large impact on the network performance. To understand typical locations and times of non-recurrent events the traffic conditions need to be evaluated per day over some time period, preferably in combination with other sources of information such as incident databases.

One might be tempted to think that traffic jams are the problem, and that the desired situation is one without them. However, by looking at the actual causes and consequences of congestion, insight is gained into control solutions that truly improve the situation. The phenomena that are generally responsible for suboptimal network outflow are:

- **Capacity drop.** Since the free-flow capacity is 10 to 15% higher than the queue discharge rate Hall & Agyemang-Duah (1991), preventing congestion will maximize the freeway throughput;
- **Blocking back of queues.** Traffic flows in the network that do not travel passed the bottleneck location should not be unnecessarily hindered by spill-back of queues over upstream-located bifurcation points;
- **Suboptimal use of available network capacity.** Underutilization of available network infrastructure results in suboptimal network outflow, i.e. not using all capacity within the routes between an origin and destination pair, or suboptimal use of intersection capacity due to inefficient green time distributions over the signal phases.

The root causes of these phenomena are aspects such as: *too high traffic inflow to the bottleneck location*, *insufficient network capacity*, *undesirable route choice of drivers* and *undesirable traffic distribution over available lanes*.

Different types of traffic data can be used to diagnose undesired network conditions. Network state estimates are typically based on traffic flow measurements gathered by means of loop detectors, cameras and floating car data. Let us briefly review some important data sources and their application areas:

- **Freeway flow, speed, density.** These fundamentally related macroscopic traffic flow characteristics are measured by means of loop detectors that are equally spaced over the freeway. These variables can be used to identify freeway conditions over time and space, and to estimate freeway capacity. This in turn, enables us to deploy controllers that aim at keeping the freeway at capacity or to estimate freeway flow breakdown probabilities given historical data on notorious bottleneck locations.

- **Queue lengths.** On the urban network the detection means are limited, i.e. traffic is predominantly measured by the induction loops located at traffic responsive intersection controllers. If there are sufficient detectors around to make a queue estimation, we can monitor the available buffer space to hold back vehicles. The size of buffers is constrained by upstream located bifurcations or maximally acceptable waiting times.

- **Travel times.** If there are limited detection loop available at the urban network, travel times can be used to assess the quality of a route and estimate vehicle loss hours. They are typically computed from the speed measurements, or measured directly based on licence plate recognition.

- **Turn fractions.** Turn fractions of traffic tell us something about route choice and the efficiency with which different infrastructure elements can be used to hold back traffic (see Chapter 4). During undersaturated conditions, these fractions might be directly measurable at the

loops located behind the stop lines of the different directions. However, during oversaturated conditions measuring the destination dependent fractions is not ambiguous, meaning that more complex estimation methods are required.

- **Network service levels.** The network production in terms of average network outflow or speed can be fundamentally related to the accumulation (density) of vehicles within the different parts of the road network. This Macroscopic Fundamental Diagram Daganzo & Geroliminis (2008); Geroliminis & Daganzo (2008) shows a maximum network production for some critical accumulation. Based on the definition of different network service production levels, decisions can be made on aspects such as control target values, applied parameters and buffer sizes. The relation can also be used to gain insight into the storage space characteristics of a complete network, i.e. how many vehicles can be allowed into the network before its production will decrease.

8.3 Potential solutions to the problems

It is important to understand under which conditions a control intervention is expected to solve the problem or at least improve the situation. As elaborately discussed in Chapter 4, the benefits of applying coordination need to be larger than the delays caused by holding back vehicles elsewhere in the network. This is directly related to aspects as the setup of the system, e.g. the buffers that are used in the coordination and decisions on when to shut down the system.

EXAMPLE: What should the system do at the moment a freeway flow breakdown? Benefits can be achieved, if enough traffic can be held back at upstream buffers to make the system recover quickly. The network performance might however also unnecessary worsen, if the system keeps holding back traffic without resolving the breakdown. The success of such actions is strongly situation dependent and therefore needs thorough evaluation before making a design decision.

To postpone undesired traffic phenomena different archetype solutions can be identified, which can be operationalized by coordinated and integrated control strategies[1].

8.3.1 Postponing breakdown and spill-back at freeway

A freeway break down can be prevented by *local and coordinated ramp metering* to limit the inflow to the bottleneck location. Its effectiveness is related to the amount of traffic that can be held back at the ramps and the spacing between the ramps[2]. Even if the system would not be able to prevent a breakdown, holding back traffic at the ramps to postpone congestion spill-back over upstream located off-ramps might result in increased network performance.

8.3.2 Postponing spill-back at off-ramp

Queue spill-back from an off-ramp to the freeway is caused by a capacity constraint at the intersection downstream of the ramp. An effective solution is to redistribute the capacity of the intersections in the urban arterial by *coordinated intersection control*. Coordination is required, because the increase flow from the off-ramp might result in problems further downstream the arterial. Coordinated ramp metering is a less effective means to stabilize a ramp queue. Freeways are normally used for long distance trips to other urban regions. The fractions of traffic leaving the freeway at other ramps will therefore be limited, making it impossible to stabilize an off-ramp queue by holding back traffic at upstream ramps.

[1] The solutions discussed below are based on the requirement that the ramps along the freeway are equipped with ramp metering and that most intersections within the urban area are controlled and vehicle responsive.

[2] The further ramps are located apart, the longer it takes for a control signal to have an impact further downstream, and the more difficult it is to prevent a breakdown.

8.3.3 Postponing spill-back within urban arterial

Queues within the urban network can cause severe hindrance to other ongoing directions, resulting in reduced network outflow. *Coordinated intersection control* in the urban arterial can be used to stabilize queues by reducing their inflow, and to maximize the arterial's throughput and improve travel time conditions. As discussed in Chapter 6, by means of *Integrated ramp metering with intersection control* at the connections, ramp saturation can be postponed in order to increase the metering duration on a freeway bottleneck.

There are other solutions that can be explored to deal with freeway traffic problems, such as the application of dynamic speed limits to solve backward propagating shockwaves in the main stream flow Hegyi & Hoogendoorn (2010) or the application of dynamic peak lanes. However, these solutions have been explored in a later stage of the project and therefor not further elaborated on in this chapter.

8.4 Preparing the infrastructure

During the implementation process the importance of having a properly functioning installed (hardware) base is important as is stressed in Hoogendoorn et al. (2016). Also the existing operational environment needs to be prepared to maximize to impact of network wide control. This means that:

- The capacity of the network is maximized by: introducing extra lane capacity between interchanges, by separating local and through going traffic, and by improving the lay out of weaving sections and interchanges;
- All relevant traffic control and monitoring systems such as ramp meters, intersection controllers and detectors operate correctly (technically and functionally);
- Sufficient detectors are present at controlled intersections to correctly estimate queue lengths;
- Local control measures are prepared for communication with higher level control units that arrange the coordination;

- New actuators (DRIPs, ramp metering, intersection controllers) are installed to communicate interventions to the road user;
- Consistent and reliable interventions and traffic information is given to ensure that the system is credible.

These lessons may seem trivial, but getting these basics up to the required level was found to be more involved than expected Hoogendoorn et al. (2016) during the field operational test. The result was that the system became fully operational just before the start of the trial period, leaving limited time available for tuning the system parameters and to optimize the system performance.

8.5 Operationalization of the system

To ready the algorithms for operationalization in practice, different types of traffic flow models and corresponding controller implementations can be useful; each evaluating different design aspects and technically correct implementation. In the end, designers need to become familiar with aspects such as: the behavior of each individual control and monitoring unit, the interaction behavior between them in the overall system framework, and a proper system setup including well tuned parameters.

Both the individual components and the system architecture evolved through a design cycle from conceptual tests in different simulation environments to technical and functional tests on the production system. The aim of the process is firstly, to come up with a final prototype code of involved algorithms. Secondly, these prototype codes serve as a blue print for building the production system by software engineers from industry. The workflow has been structured by five typical phases, being:

- Conceptual design and tests by means of simulation models;
- Building the production system;
- Technical tests for validating correct implementation of the operational algorithms;

- Tuning of involved parameters to maximize the production system's performance;

- Functional tests for validating the system behavior in practice under typical circumstances.

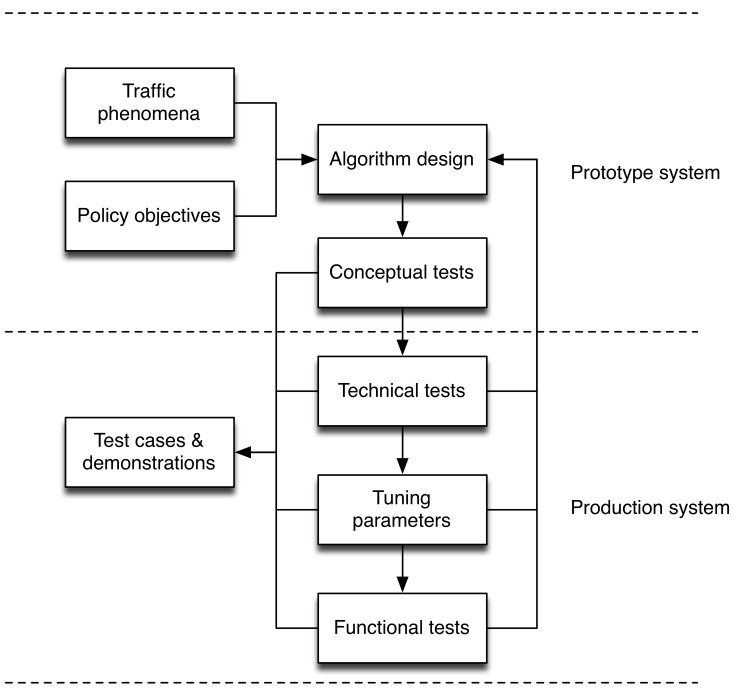

Figure 8.1: Overview design phases of the system within the Field Operational Test Amsterdam.

In this section we will briefly elaborate on these key design phases that stimulate a stepwise development process in which the developers maintain in charge of the design. In the end, successful operationalization is fully dependent on thorough understanding of the individual components and the overall system behavior.

8.5.1 Conceptual designs

The conceptual designs and tests are the basis for studying, validating and improving a new algorithm. Simulation models are used to test control and monitoring concepts' basic functionality. This implies that the algorithm is designed based on a thorough understanding of the problem at hand and its foreseen solution. In the context of the development of integrated network management both the individual components as well as the overall framework is conceptually tested in a variety of simulation environments. The following recommendations are made for setting up conceptual tests:

Work from simple to more complex.

Initially many simplifications are helpful to keep the controller behavior comprehensible for unambiguous evaluation of the control strategies. Implementing the controller in a minimum, but non-trivial network, makes it easy to understand the basic principles on which the strategy is based. Hence, it is recommended to limit the number of controllers involved and to use simple demand profiles (easy numbers) to completely orchestrate the test scenarios.

Test hypothesis on desired outcomes.

Given the network demand and supply characteristics that form a scenario, clear a priori hypotheses on the expected outcomes are tested to see if controllers or monitoring units work in line with our expectations. Note that this is fundamentally different from running a random scenario, evaluating the outcomes to see if the controller behavior is more or less reasonable and if the network performance is improved.

Gradually introduce stochasticity and more complexity.

To already know what a controller does in a simple not-trivial set-up under ideal conditions, makes interpreting control signals and network state estimates much easier when complexity is increased. Increasing complexity is done by increasing the number of applied actuators and by introducing more stochasticity on the traffic demands to explore if the controller properly deals

with demand fluctuations and uncertainties. Moreover, other aspects to explore are delayed control inputs or different data aggregation levels.

Use a suitable model

The only requirement for a simulation environment is that it captures all essential phenomena. For instance, when testing how the controller fills storage spaces, there is no need to put a lot of effort in replicating the capacity drop correctly. However, when we want to get a first feeling on the control benefits, the model needs to correctly represent the phenomena that decrease the network performance. As complexity of the evaluations increase, it might be helpful to implement the controller in more advanced modeling environments, with more advanced representation of the traffic flow dynamics and traffic phenomena.

Build comprehensible test cases

Comprehensible test cases are important to demonstrate the potential of a concept to fellow project participants, contractors, other stakeholders and even outsiders to gain support for the innovative ideas. Moreover, gaining momentum, belief and enthusiasm are important factors that lead to success. Hence, test cases need to have a clear story about the typical problem that is encountered, how the scenario is build, and what the controllers will do solve the problem, and finally to give insight into the benefits of the approach.

Test of the shelf concepts

It can be beneficial and cost efficient to adopt existing concepts for coordination, which have already been operational in practice elsewhere. These of-the-shelf products normally need to be adjusted to be fully compatible in the envisioned overall framework. This implies, that there are conceptual changes required to a product that is already operationalized. As goes for all new concepts, it is highly recommendable that these changes are conceptually tested, before being implemented in the complex setting of an integrated network wide control framework.

WORKED EXAMPLE: Coordinated ramp metering algorithm

In this section we will briefly illustrate the different model implementations that were realized for the conceptual tests on the designed *coordinated ramp metering algorithm*. We will explain by means of a worked example how the components within the Field Operational Test Amsterdam were prepared for operationalization in a production environment. It is shown how the algorithm has been implemented in a variety of simulation environments, ranging from minimum, but non-trivial, to a large-scale implementation based on a calibrated network of the Amsterdam region. Obviously, making multiple simulation implementations of the algorithm takes time, but the time savings that are realized during the actual operationalization in the production environment of the system are probably larger due to the gained in-depth understanding of the algorithm.

- **Conceptual testing - Testing basic algorithm functionality.** Initially a very elementary 5-link implementation (3 freeway links, 2 ramps) was realized in a store-and-forward model to test the coordination mechanism. In Figure 8.2 its basic functioning can be seen by looking at the saturation degrees of the ramps and the saturation duration of the master ramp. Typical characteristics of store-and-forward models are the simple propagation of flow per simulation time step to downstream links, free flow travel times over the link length and the formation of vertical queues at the downstream ends of the links in case of oversaturated conditions. This type of modeling is well suited to evaluate the workings of the controller with respect to the use of storage space, however, traffic dynamics within a link are less realistic.

- **Conceptual testing - Increase complexity.** In order to better understand and evaluate the resulting system dynamics, an implementation is made in a first-order cell transmission traffic flow model that is able to realistically simulate the impact of queue spill -back and the capacity drop. Because of the more realistic representation of traffic flow dynamics, we can also check the impact of the controller and its setup on the network state for minimum, not-trivial 2 or 3 ramp networks with comprehensible demand profiles. In this stage, initial evaluations

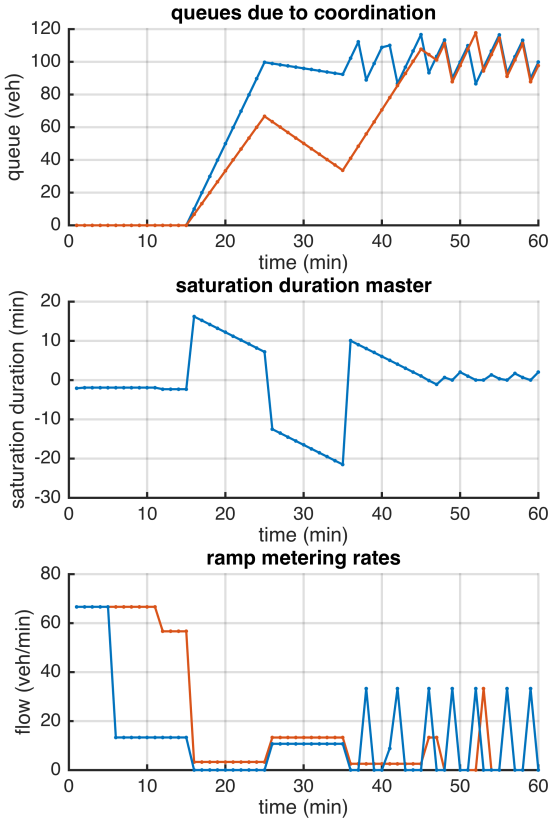

Figure 8.2: Conceptual test results for the coordinated ramp metering algorithm using a store-and-forward modeling approach. The blue line indicates the master ramp and the red line the slave ramp. As the master ramp is holding back traffic, its saturation duration is positive, leading to the incentive for the assisting slave ramps to also hold back traffic. As the gaps arrive at the master, it can release more traffic, resulting in a negative saturation duration (indicating its queue is dissolving) and the incentive for the slave ramp to also release more traffic. As the corresponding higher flow regime arrives at the master, it starts holding back traffic again, resulting in positive saturation duration.

can be made on the controller with respect to stochastic and highly fluctuating demands and different ramp characteristics.

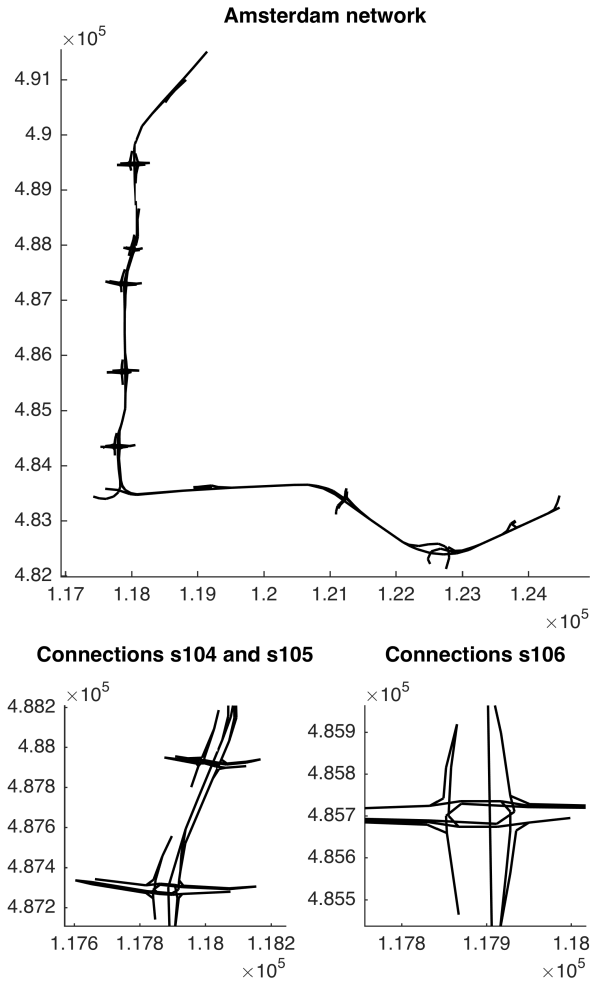

Figure 8.3: Network layout of the full calibrated model of Amsterdam region in macroscopic modeling environment.

- **Conceptual testing - Implementation on network level.** Once a good understanding is gained and necessary adjustments to the algorithm made, then we can gradually scale up to a full-scale calibrated model of the Amsterdam region as shown in Figure 8.3. This enables

us to explore the potential of the control approach given typical peak period conditions in the region. Moreover, it also allows for evaluating aspects such as a proper setup of the system (e.g. master, slave configurations as discussed in Section 7.2.10) and the applicability of the approach under more realistic conditions.

As will be clear from Figure 8.4a, b,c, proper functioning of the controller is much easier analysed for a simple test case than for a scenario including stochasticity, or the complexity of true network layouts and demand patterns. However, knowing the expected control and network dynamics, makes the results in case of random arrival patterns, complex demand patterns, and the large-scale Amsterdam network very plausible.

As a final conceptual step, the coordination algorithms have been implemented in the microscopic simulation environment Vissim, based on a Matlab interface. The focus was put on testing the interfaces within the system framework, including the more realistic controllers such as the embedded ramp metering installations and intersection controllers in the Vissim software. This prototype environment was also used to design and test the data flows within the system and to fine-tune the input and output relationships between the units. Examples and outcomes from this conceptual test phase can be found in chapter 7.

8.5.2 Mock-up of production system

Once the prototype algorithms for the control and monitoring units are thoroughly understood and tested, and when they correctly interact, than the system needs to be build, setup and tested in an environment that can replicate many of the limitations and unruliness of the situation in the field. The focus during technical tests is put on how the control and monitoring units are interacting with real operational actuators and a highly stochastic traffic flow process.

To this aim, a mock-up of the production system was built that would initially interact with a *simulation environment* identical to reality and later brought on-line in practice. Different parties from industry translated the

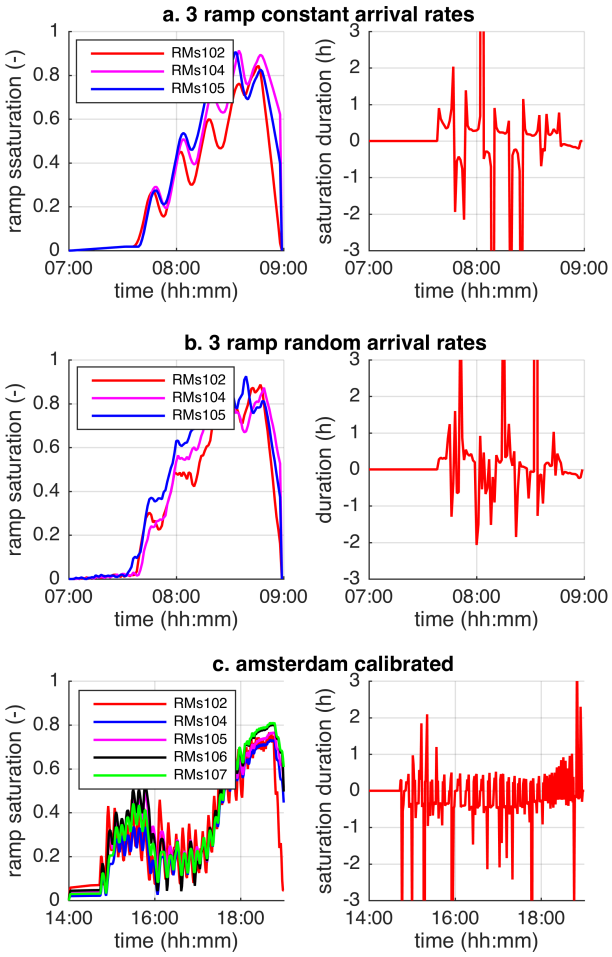

Figure 8.4: Conceptual test results of the coordinated ramp metering algorithm using a macroscopic first-order simulation environment with (a) a simple network layout and deterministic demands, (b) a simple network layout and stochastic demands and (c) large-scale implementation with realistic demands.

prototype codes into applications for the production system. The difficulty becoming that different industrial partners are now responsible to keep the system operational, their codes bug-free and their servers in the air. Note that this simulation environment embedded:

- Real intersection control and ramp metering software from industrial parties;
- Real interfaces for deploying the different actuators;
- Real monitoring devices such as freeway and urban loop detectors for measuring traffic flow and queuing characteristics.

8.5.3 Technical testing

It is relatively easy to debug the production system, as long as the knowledge is preserved from the conceptual testing phase. The engineers testing proper functioning of the production system are not necessarily the same engineers that designed the algorithms, meaning that valuable knowledge on all concepts could be lost. The advantages of using a mock-up are:

- The system can be thoroughly tested and debugged before going online;
- This minimizes unnecessary hindrance to traffic;
- It prevents unsafe situations due to system errors as much as possible;
- The ground truth system states are available in the mock-up, so that the monitoring units can actually be debugged.

A lot of data circulates within the system and proper error handling is required in case components fail or communication is lost. The impact of the following aspects are useful to test in the production environment with a simulation interface:

- **Monitoring delays.** It takes time in reality to publish the monitored data for the control units, meaning that control actions are not based on the actual conditions in the network.

- **Data aggregation levels.** Control signals are determined for time intervals that are typically larger than the data monitoring intervals. It is important to evaluate if the monitored state is stable enough to use the most recent data samples, or if (and to what degree) data needs to be aggregated.

- **Monitoring quality.** It needs consideration to what extent the system is able to make good decisions based on the input data. In a realistic environment, many control inputs are based on state estimations such as queue lengths which are based on occupancy measures of a limited set of loopdetectors.

- **Interfacing.** Realizing coordination between measures requires preparing existing actuators – which are operating predominantly on a local level – for deploying control from a network perspective.

- **Time synchronization.** Each monitoring and control unit has its own timing. In case multiple monitoring and control units need to collaborate, their activities need synchronization to enable the use of up-to-date information.

- **Data storage.** It is important that the data is stored such that it enables debugging of the individual components and the complete system architecture and to evaluate the impact on the network performance.

- **Simulation setup.** It is wise to prepare the setup of the final simulation environment –in which the system is technically and functionally tested–, such that potential technical and operational issues can be evaluated. Note that operationalization in practice is much more cumbersome than operationalization in a central simulation environment, due to aspects as communication issues and delays, technicalities, unit failures and all kinds of errors.

8.5.4 Tuning of parameters

Controllers and monitoring units contain parameters that can be to change the behavior of the controller. These parameters can be determined based on theoretical considerations, expert knowledge or trial-and-error procedures. It is required that the designer understands the relation between the controller

behavior (performance) and the controller parameters. If the traffic conditions change over time, these parameters need re-tuning.

Simulation environments are a very suitable place to gain experience with the effects of changing parameters. However, keep in mind that parameters –set during simulation with the mock-up system– need re-tuning during the employment in the field. It is simply impossible to accurately mimic the prevailing conditions in practice within a simulation model. Simulation therefore serves the purpose of making a reasonable first setting of parameters and gaining experience in finding good settings as fast as possible.

Trial-and-error tuning takes a lot of time; especially when tuning is done in large-scale simulation models or during real-time operation. Systematic tuning approaches are in that respect desirable and potentially save a lot of time. In Chapter 6 an example of such systematic tuning approach is given, which has been specifically tailored to the needs within the Field Operational Test Amsterdam.

8.5.5 Functional testing

If the system framework and its components are properly implemented and set up, the functional tests can be executed on the production system. In first instance, this is done based on simulation tests with the the mock-up of the system. This means that the system is fed different scenarios that provoke control actions that can be tested.

Hypotheses are identified that reflect the expected outcomes and desired system behavior. The simulation results are then analyzed and validated step-by-step based on raw and processed monitoring data, control inputs, control variables, control outputs and resulting freeway and urban states. It can be helpful to build a story on what the controller does during control intervals at crucial moments. Hypotheses can then be checked based on the raw monitoring data, the processed data, up to working mechanisms of involved controllers. Testing hypotheses in a production environment (build and maintained by one or more parties from industry) nevertheless remains complex due to the fact that:

- The system is operating on large-scale and fully setup, involving all components and many parameter settings;
- During the evaluation of the system limited means are available to validate the system in real-time;
- There is no direct access to the actual software to evaluate (i.e. under the hood) the components' functionality;
- System output is limited to real-time data that circulates in the system, i.e. making it impossible to look at time-series for debugging purposes;
- Debugging is based on post processing the historical data logs in the system;
- Involved software engineers (that build the system) and the traffic engineers (that designed the system) are not in one room during these tests, hence the process of recognizing, diagnosing and solving problems takes time.

In Figure 8.5 an example is given of the application developed for displaying the control and monitoring outputs of the system in the production environment. As can be seen, the instantaneous system outputs are presented and time-series of the variables are not available to gain quick insight into the system behavior of time.

The post-processing of the data for evaluation can be a cumbersome task. During the field operational test the following steps needed to be followed before insight was gained in the system data. User interfaces were developed to download the data from a dedicated server using a VPN connection, and to parse the data into formats that allow evaluation of its content (e.g. the Matlab (.mat) format). Dedicated analysis tools were made to graphically display the monitoring and control variables over time. An overview of this procedure is given in Figure 8.6, showing that the units logged their actions through the system's communication line (PPA-Bus) in Json format.

After the functional tests, the system is ready to go live in the field. The process of technical and functional testing of the production system in practice, is identical to the tests performed with the mock-up system that interfaced with the simulation environment. However, one large disadvantage

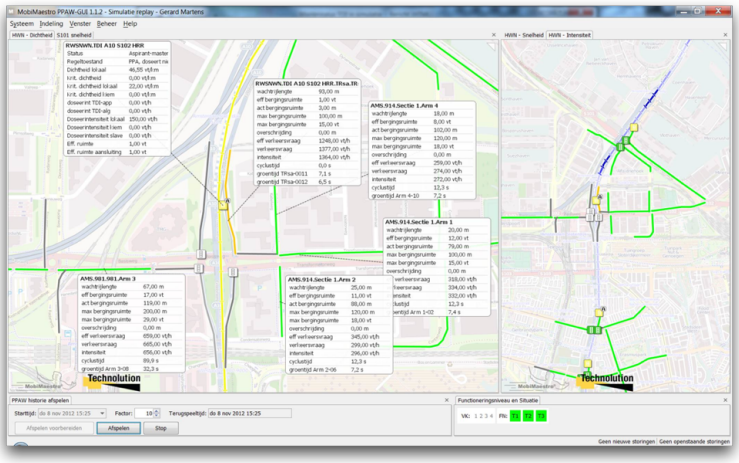

Figure 8.5: Overview of the interface developed for the production environment to display instantaneous monitoring and control variables for validation purposes on the system's functioning.

has to be taken into account: In reality there is no unambiguous ground truth network state available anymore. This implies that validation of the network states has to be done by means of remotely controlled camera's as can be seen in Figure 8.7. This is obviously less trustworthy than state descriptions retrieved from a traffic simulation model.

8.6 Conclusions

We have elaborated on the complex task of designing an integrated network management system supported by conceptual, technical and functional testing on the involved monitoring and control units and on the overall system framework. Control concepts need to be designed based on a thorough understanding of the problem at hand and a clear understanding on potential solution directions. To keep the design process towards the operational system (production system) comprehensible and manageable, we have explicitly defined the different development phases including the corresponding

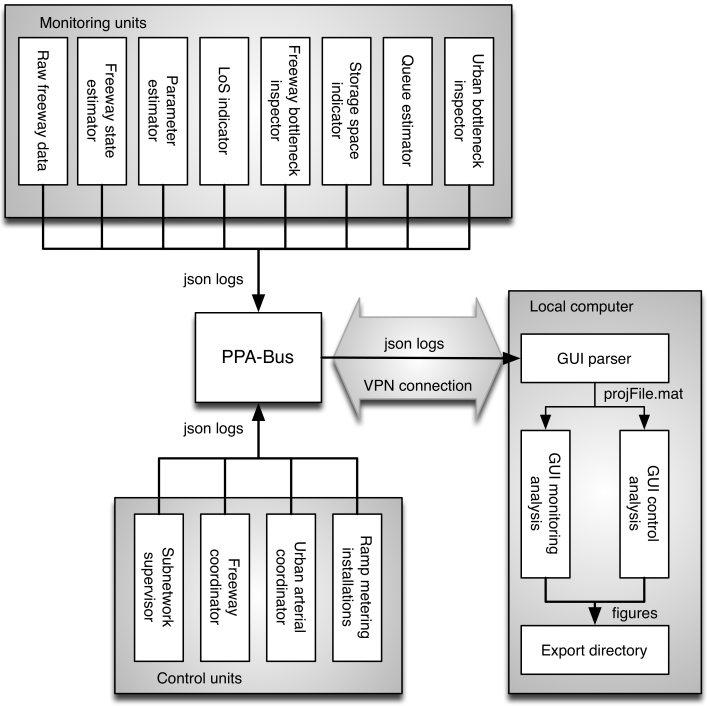

Figure 8.6: Overview interfacing to review historical data of the system operations for evaluation purposes.

Figure 8.7: Overview of cameras that were used to determine the actual state in the network, i.e. to check upon the proper functioning of monitoring units as the Freeway State Estimator and the Queue Estimator.

considerations that lead to a good design, technically correct implementation and desired functional behavior.

Many different models are used by the traffic engineers that design the system and prepare prototype codes of the system components. As we have seen, each model is used to zoom in at specific design aspects that need testing. The codes are subsequently used by engineers from industry to build the production system, as well as a mock-up of this system that connects to a simulation environment for debugging and testing purposes.

The complexity during tests is gradually increased in terms of the used type of model, the network and the controller implementation. It can be recommended to start with a minimum, but non-trivial network and control implementation to unambiguously validate the concepts. Only then more complex conditions can be evaluated, such as stochastic arrival rates, realistic demand patterns and large-scale network layouts.

The hypotheses that are used to validate the concept, trace technical errors and undesired functionality should reflect the expected outcomes and desired behavior of the system. This makes simulation a check on the functionality of the system, based on thorough understanding of its workings and its foreseen impact on the network state. To conclude, based on the experience within the Field Operational Test Amsterdam some final remarks:

- Despite time pressure, sufficient time needs to be allocated for the tuning of system parameters and the validation of the functional behavior of the system;
- The control units should determine the required data quality and time intervals of monitoring units, not vice versa;
- For real-time validation of the system time series of available data are needed, and not only instant real-time values of parameters and system variables;
- Comprehensible test cases that clearly illustrate the potential of the system, help gaining support and engagement for integrated network management in practice;
- Operationalizing the system and scaling up is done step wise to keep track of what is happening.

Chapter 9

Conclusions and recommendations

This dissertation originated from the ambition to develop new control methods for traffic management that focus on aspects such as solving problems from a network perspective, integrating traffic measures, policy objectives, user interests and practical applicability.

The following two general needs for the set-up and implementation of advanced network-wide traffic management systems can be drawn from the research results:

- In practice, there is a need for an integrated, network-wide control system that can combine state-of-the-art measurement and control methods to identify traffic problems promptly and tackle the root causes;

- Because of their characteristic vulnerabilities, currently available control methods can rarely be deployed effectively and efficiently on a large scale (e.g. for optimizing control strategies and integrated control scenarios). Theoretical and operational complexity make these approaches error-prone, expensive to maintain, difficult to operate and hard to apply in real-time.

This dissertation develops a framework that efficiently operationalizes integrated (coordinated, network-wide) traffic management, along with an elaboration of effective measuring and controlling strategies with the associ-

ated algorithms and overall design approach. The overall conclusions stemming from our research are as follows:

1. **In order to realize a technically feasible integrated control system that is sufficiently simplified in practice to be applicable, in-depth analysis is needed to get to grips with the root causes of the traffic problems in the network in question.** This implies that the developers are able to take stock of the situation and get to the bottom of the complexity of the traffic, control technology and policy-related aspects.

2. **The road infrastructure and the traffic management areal must be improved and updated before maximal impact can be realized with network-wide traffic management:**

   - In addition, it must be possible to measure and control the essential roads, which, in all likelihood, requires the expansion of the number of traffic lights, ramp meters and dynamic route information panels (DRIPs);
   - Finally, local actuators will have to be modified for supra-local and integrated management by coordinated control strategies that solve traffic problems based on a network perspective.

3. **The following matters require particular attention when designing network-wide control strategies:**

   - Controllers must be able to rapidly realize the desired stable state, on the one hand, and to deal with unforeseen changes in the traffic demand and infrastructure capacity, on the other;
   - The behavior of controllers must be comprehensible so that they can be tested to ensure that they are operating correctly and so that what is happening in the network can be explained;
   - If the system is to be rolled out on a large scale, it is important that the computational demand remains limited for real-time application;
   - Last but not least, in the event of intervention, the travel time gains must always be greater than the delays incurred by buffering traffic elsewhere in the network.

These conclusions are explained below in line with the research questions posed in the introduction section of the thesis. Firstly, Section 9.1 looks at the answers to the research questions that led to the development of a technically practicable approach for an integrated, network-wide control system. Section 9.2 continues with an overview of recommendations for policy and practice, based on experience gained from the Field Operational Test Amsterdam (PPA), a large-scale pilot for integrated traffic management in Amsterdam. Section 9.3 concludes with recommendations for future research and the further development of the network-wide traffic management approach presented here.

9.1 Research questions and conclusions

This section elaborates on the answers to the research questions that led to the development of a technically practicable approach for an integrated, network-wide control system.

9.1.1 What traffic phenomena affect the performance of a traffic network?

The study of the literature on traffic control engineering described in Chapter 2 revealed that the occurrence of a limited number of phenomena must be prevented or postponed to improve the performance of a traffic network; these are as follows:

- Capacity drop on the freeway after congestion has arisen;
- Spill back (blocking back) of congestion or queues over upstream forks or intersections;
- Suboptimal route and lane choices of vehicles and under-utilization of network capacity (e.g. at intersections).

Test cases are used to illustrate how the control methods proposed in this dissertation can reduce the impact of the aforementioned phenomena. The chosen control objective obviously affects the control strategy used to lead to optimality. For example: Minimizing the total time spent within the network, while the value per time unit of a vehicle is incorporated (private cars,

trucks, public transport buses full of people, etc.), may require a different intervention than if all vehicles are deemed equivalent to one another.

Besides the limited number of phenomena that affect the network performance, there are also a limited number of fundamental causes and therefore possible solutions for these phenomena. Recurrent congestion and queues are caused by an imbalance between available network capacity and the traffic demand in the direction of the bottleneck. An intervention therefore always affects the traffic demand in the direction of the bottleneck or the network capacity in the downstream direction.

9.1.2 How can the impact of these phenomena be limited?

The traffic demand in the direction of the bottleneck and network capacity can be influenced by an enormous diversity of measures. These include intersection control systems, ramp metering, dynamic speed regimes and DRIPs. Such measures, based on local and coordinated control methods, can reroute traffic over alternative parts of the network and hold back or allow traffic through to specific network locations. The more space and route capacity there is available to temporarily buffer or divert traffic, the longer spill back and capacity drop can be prevented.

When considering the literature on control strategies and methods in Chapter 2, we ascertained that some of the proposed contributions combat symptoms, rather than tackling the causes, which can lead to suboptimal, or even worse, network conditions. Generally speaking, there is plenty of room for improvement in the most recent developments. First and foremost: A control system must tackle the underlying causes of undesirable phenomena in an explainable manner. If a system does not, it should not be put into practical operation.

Furthermore, it is crucial to realize that the coordination of measures does not, by definition, lead to improvement in network performance. After all, in order to prevent delays at a potential bottleneck, the traffic must be held back or diverted to longer routes elsewhere in the network. The effectiveness of the coordination will then be determined by:

- **The ultimate relationship between travel time gains and losses.** It is possible that the collective delay in the buffers is greater than the travel time saved in the bottleneck and, in this case, coordination is not worthwhile;

- **The capacity of the control strategy to utilize the available network capacity to the full** in terms of the efficient use of buffer space and road capacity.

9.1.3 How can the most recent control strategies and methods be improved?

Based on the aforementioned points, this research proposes a comprehensible and practicable improvement for the following control strategies and methods:

Service level-based routing. The discussion of the 'state of the art' in Chapter 2 shows that, in practice, traffic is routed to achieve user equilibrium conditions or, in the event of an emergency, to an alternative route. There are no operational routing strategies available in practice that focus on improving network performance while reasonably safeguarding the interests of road users.

The routing strategy proposed in Chapter 3 maximizes utilization of route capacity and limits the development of undesirable phenomena such as spill back and capacity drop. Excessive delays for rerouted road users are limited because the controllers observe maximal travel time differences between the routes.

- The service levels of the routes are reduced and increased, in steps, to realize network conditions in line with policy objectives.

- The method is able to fully utilize the available capacity over the regulated routes and limit spillback effects - including when multiple controllers are active on overlapping routes or parts of routes.

- If travel time on a route is affected by multiple bottlenecks, it is not possible to adequately prevent spill back by enforcing a certain service level. However, stepwise reduction in the service level will help delay spill back of congestion or the moment the freeway flow breaks down (with the associated capacity drop).

Coordinated control on (and to) freeways. The effectiveness of coordinated ramp metering strategies depends on the degree to which the freeway capacity and the space on the coordinated ramps are utilized. If queues develop on the freeway and there is still space available on the on-ramps or in the traffic flow (towards the bottleneck), control is suboptimal.

The ramp metering strategy proposed in Chapter 5 maximizes utilization of this freeway capacity and buffer space by allowing the coordinated on-ramps to fill up in the downstream order. This gives the space realized in the mainstream traffic flow at upstream coordinated on-ramps the time to be utilized at the location of the bottleneck in question.

- The approach is based on the synchronization of on-ramp saturation times;
- It has been shown that allowing on-ramps to fill up in the downstream order leads to an improvement in network performance;
- If there are structural measuring faults in the available space on, and traffic demand towards, the on-ramps, this can lead to on-ramps not becoming saturated in the correct order. This again results in suboptimal network performance.

Integrated control on junctions and coordinated control on urban axes. If traffic is to be effectively buffered at intersections, it is essential that the coordinated or integrated approaches can override the operation of existing vehicle-dependent local and coordinated intersection controls in a simple manner. This makes it possible to:

- Delay capacity drop on the freeway by holding back traffic on the junction upstream of an on-ramp;
- Limit negative impact of spill back of queues on the underlying road network by holding back traffic at more favorable upstream locations (intersections).

Section 6.2 presents a control strategy that prevents spill back of queues on the underlying road network by buffering traffic at intersections located upstream. The approach fills all the buffers simultaneously (in parallel) based on the available space in the critical link with the bottleneck. The network continues to perform well because hindrance to through traffic flows is limited.

- The approach is illustrated for the coordination of intersections on an urban axis and can be made even more effective by looking at the order in which the buffers fill up. After all, through traffic that does not need to pass through the bottleneck, is hindered in all the buffers simultaneously from the moment the coordination is activated.

To deal with this, a strategy is introduced in Section 6.3 which minimizes inconvenience to through traffic by filling buffers successively (in series) in the order of decreasing traffic fraction in the direction of the bottleneck.

- The approach is elaborated for the integrated control of ramp metering and the intersection located immediately upstream of the on-ramp in question. By holding back traffic at the upstream intersection, the on-ramp can meter for a longer period on a freeway bottleneck. Prolonged prevention of capacity drop on the freeway in the rush hour yields a substantial travel time gain, as is also illustrated in Section 4.5.

- To be able to hold back enough traffic to stabilize a queue, clusters of cooperating buffers may have to be defined. A suitable approach to doing so is also presented in Section 6.3.

- The test results show that the filling of buffers in series is favorable if the traffic fractions in the direction of the bottleneck of the buffers in question vary greatly.

Systematic tuning of parameters. All of the aforementioned controllers are of the 'feedback' type. As such, they can realize the desired state in the network and deal with unforeseen circumstances. It is important that feedback controllers can be rapidly and correctly set in an operational environment because their parameters (i.e. feedback gains) determine how adequately they function. Furthermore, we soon end up having to deal with dozens of feedback controllers that all have their own optimal settings. One might

wonder whether it is possible to determine optimal settings using trial-and-error methods.

Section 6.5 presents a method for tuning the gains systematically. Using a state-space model of the queue dynamics with the controllers, it is possible to identify the theoretically optimal setting. This might need fine tuning in a practical environment, but this at least gives an idea of the order of magnitude of the values required for effective control. Much time can therefore be saved, and the effectiveness of the controllers increased, using this approach.

9.1.4 How must buffers be selected and deployed to maximize the improvement of the network performance?

Chapter 4 uses cumulative curves to show the costs and benefits of the coordinated deployment of the various measures. This method can be used to arrive at substantiated options for the deployment of buffers. The application of the method has been demonstrated for coordinated ramp metering and integrated control of an on-ramp and upstream intersection.

It is vital that the set of buffers for coordination are chosen carefully. The number and deployment capacity of the buffers required to achieve maximal improvement of the network performance are situation specific. Selection depends on factors such as the duration and degree of overload of the bottleneck, the capacity drop and the traffic fraction in the buffer travelling in the direction of the bottleneck.

- The higher the potential delay in the bottleneck, the faster a buffer is allowed into the selection or, to put it another way: the smaller the fraction of traffic travelling in the direction of the bottleneck has to be to effectively include the buffer in the coordination.

- The buffering of vehicles can also have a negative effect on the network performance. This is the case if too many vehicles are held back and the release of these vehicles causes the bottleneck to be overloaded for a prolonged period.

Section 4.8 looks at the impact on the network performance of different strategies for allowing buffers to fill up. If there are big differences in the fractions travelling in the direction of the bottleneck, it is often worthwhile to fill buffers in series. If the differences are slight and the bottleneck only briefly overloaded, it is better to fill the buffers in parallel (simultaneously). The variant that works most effectively must be investigated in advance on the basis of the typical network conditions.

Analyzing the costs and benefits of coordination using parameterized cumulative curves is an accessible way of making coordinated control strategies more effective. The use of this type of graphic approach may result in more support in the integrated development and rolling out of coordinated and network-wide traffic management in operational practice. Such instruments provide points of reference for analyzing the costs and benefits and configuring control strategies.

9.2 Recommendations for policy and practice

This section gives an overview of recommendations for policy and practice, based on experience gained from the Field Operational Test Amsterdam (PPA), a large-scale pilot for integrated traffic management in Amsterdam.

9.2.1 How can traffic control strategies operationalize road authorities' policy objectives?

In an integrated approach, traffic management must be put into the context of operational practice: an area not only has to be accessible, but also livable. Policy objectives like this can only be put into practice if operational measurement and control systems enable the translation of these objectives into the desired traffic conditions in the network.

To this end, control strategies must aim at specific conditions, the parameter settings determining how the strategies put the policy objectives into practice. The pursuit of policy objectives, such as livability, has a direct effect on the degree to which network performance can be maximized.

- Chapters 3 shows that when redistributing traffic over route alternatives, equity can be guaranteed by observing a maximal travel time difference.

- Chapter 7 explains that buffers' size can be chosen depending on the situation and that this is a function of the maximal acceptable delay (in one or more buffers connected in series) or maximal acceptable queue length in an environment.

- Chapter 5 discusses various ramp metering algorithms that can target specific road intensities or densities. These targets can also represent policy objectives relating to air quality (emission per vehicle kilometer), safety (vehicle kilometers traveled) or living enjoyment (maximal intensity and queues).

9.2.2 What preconditions does operational practice impose on coordinated network-wide traffic management?

Coordinated control strategies and the associated algorithms have to meet certain preconditions to be applicable in the existing operational practice:

- **Comprehensibility.** It is essential that a developer is able to test the correct functional behavior of concepts and technically-correct implementation under typical traffic conditions. Those responsible for policy must also be able to account for the impact on the traffic situation in the network (free-flowing traffic on the freeway vs. full buffers in the city) to the general public and government.

  In more complex cases, at the operational level it may be difficult to figure out precisely what a controller does to improve network performance. This is not a problem as long as undesirable phenomena are adequately prevented or solved and no apparently illogical tactical or strategic situations arise.

As already mentioned, all the dynamic control strategies proposed in Chapters 3, 5 and 6 realize straightforward interventions with the associated explainable network conditions.

- **Scalability.** The proposed control strategies are more easily scalable than optimizing control strategies at network level because they are reactive in nature and based on domain knowledge. After all, the computational demand remains limited if control signals are only based on a current and a desired (measured or predicted) network state. This is quite different from the theoretical optimizing controls that determine the control signal on the basis of iterative predictions and are therefore difficult to apply on a large scale in real-time.

Section 3.6 compares an optimizing (MPC-based) controller with the proposed controller for routing traffic based on service levels. The test case results show that, in this example, the computational demand of the proposed controller scales linearly and that of the MPC controller, exponentially (with the number of variables to be optimized).

- **Ability to cope with unforeseen situations.** The operational practice is unpredictable; unforeseen incidents that have an impact on the traffic demand or network capacity occur regularly. Nevertheless, controllers must at all times determine meaningful control signals that steer the system to the required state.

  All of the controllers proposed in this dissertation are of the feedback type to make allowances for unforeseen matters by basing the control signal on the current and desired network state. The control parameters are set such that interventions rapidly achieve the desired and stable network state.

Optimizing control strategies can also teach us a great deal about undesirable traffic phenomena. They can show optimal behavior without domain knowledge, which, in a simple test case that has been properly set up, can be interpreted clearly. However, even with recent developments which reduce the computational demand - by means of efficient parameterization, linearization of models and the hierarchical or cooperative setting up of the MPC controller - the application of optimizing strategies continues to be a challenge in large-scale traffic networks in real-time.

Moreover, the interpretation of the behavior of these controllers (combined with the resulting network conditions) requires expertise in control engineering, traffic modeling, traffic phenomena and perhaps also model

artefacts. This knowledge is scarce in practice and there are therefore few, if any, professionals available to adequately implement strategies of this kind and to maintain them on a large scale.

Resume: this dissertation puts forward control strategies which, on the basis of domain knowledge, come close to achieving system optimality and which can be modified for application in more complex network configurations.

9.2.3 How can generic methods for data collection and traffic management be integrated at network level?

Effective control signals can only be determined if adequate measurements and estimates of the state of the network are available. For control at network level, a broad arsenal of instruments is needed for both the arterial road network and the underlying roads; these include:

- Actuators: traffic lights, ramp metering, DRIPs and variable message signs;
- Control strategies and methods: local, coordinated and integrated;
- Data collection and processing methods: loops, queue length estimations, state estimation techniques or filters and the macroscopic fundamental diagram (MFD) to describe the relationship between the traffic flux, traffic density and velocity in the network.

These measurement and control components must subsequently be combined in a framework that effectively tackles problems in a random network. Chapter 7 presents the framework realized in the Field Operational Test Amsterdam. This was a large-scale pilot project which yielded critical experience with integrated traffic management as regards examining potential and acquiring a better understanding of the associated challenges in a 'learning by doing' setting.

The design of the framework presented is based on the paradigm: *"Solve problems locally if possible, more globally if needed"*. If traffic problems cannot be solved using local measures, coordinated measures must be applied. Chapter 7 presents the control strategies discussed earlier in the con-

text of the framework and Chapters 5 and 6 describe and test their algorithms extensively.

It should be noted that these controllers can only determine worthwhile interventions based on correct measurements or estimates of network conditions. Important aspects of the data collection and processing are as follows:

- Ensure that there is adequate information and analyze the state of the entire freeway (i.e. not just at detector locations). The state of the freeway can be described by the parameters: intensity, speed and density. For example, the tail of a queue can be followed and metering is possible at all locations where capacity is in danger of being exceeded.

- Control proactively by anticipating flow breakdowns. During the field operational test, the first steps were taken in estimating the likelihood of a flow breakdown and proactive control of such situations.

- Capacity estimation is a dynamic process. Capacity must continually be adjusted to allow for all the factors that affect said capacity and thus enable full utilization of the road.

- Determine the current queues as accurately as possible. With the limited space on the underlying road network and the strongly-fluctuating behavior of queues, over and underestimates can lead to suboptimal use of buffer space and inconvenience to other traffic by, for example, spill back, respectively.

- The average service level of a network can be used to make strategic decisions on the prioritization of certain archetypal situations and the allocation of buffer space for the various control objectives.

The framework has a modular structure which allows for stepwise development and implementation and to which new state-of-the-art measurement and control systems can be added. The operation of the system and its components are illustrated in detail. The results of a simulation experiment with working operational prototype code shows that the setup presented improves network conditions, provided the system is correctly setup.

9.2.4 What are the lessons learned in respect to the design and implementation of the large-scale field operational test for traffic management in Amsterdam?

Chapter 8 discusses how we can arrive at a workable and effective network-wide control system that is simplified enough in practice to be applicable. To this end, those involved must comprehend the complexity of the traffic, the control technology and the policy-related aspects. The Field Operational Test Amsterdam yielded experience in these areas and indicated that the following aspects are essential:

Preparation of existing road and control infrastructure:

- Identify the causes of the prevailing traffic problems in the network based on an traffic analysis;
- Determine the correct choice of location and scope for the system for a worthwhile and effective investment in a traffic management arsenal;
- Before applying traffic management, it is advisable to prepare the available infrastructure with dynamic lanes and by disentangling traffic flows and improving the layout of weaving sections, for example;
- Prepare local controllers for communication from above by means of coordinated controllers and install additional actuators (DRIPs, ramp meters, intersection controls) to realize total control of the network;
- Install sufficient equipment (e.g. detectors) in the underlying road network and freeway network to be able to make adequate estimates of the conditions.

Design process of the control system:

- The concepts developed must have proven capability in terms of being able to tackle traffic problems. If a controller does not show behavior that can be explained, it is not advisable to implement it in practice - this also applies to existing off-the-shelf applications that are available on the market;

- Test concepts first in a minimal non-trivial environment and gradually add more complexity (e.g. type model, system and process characteristics, stochasticity, etc.);
- Allow the controllers to determine the required quality of the measurements to be able to arrive at good control signals rather than the reverse;
- It is important to be aware of the operation and effect of the data processing techniques used. For example: Unlike Kalman filters, ASM filters for estimating conditions on freeways smoothen the measurements. This can result in under or overestimation of the true conditions with all the attendant control problems;
- For a clear description of the conditions on a freeway, there must be a dynamic overview of which detection loops are in operation and which lanes are open;
- Buffers for temporarily holding back traffic must be large enough to prevent unstable conditions (i.e. caused by ad hoc activation of queue protection strategies and large fluctuations in the control signals).

Practical implementation including tests and calibration:

- Start by testing a production environment in a mock-up of the actual situation, without obstacles, to trace any faults in the design and its technical implementation.
- Implement the system in steps to enable unambiguous testing of the components and take all the time needed for the setup and calibration;
- Determine the buffers to be used based on situation-specific analysis to guarantee that the effect of the coordination is positive;
- Make available time series of system variables during the operation for debugging purposes (settings, monitoring and control);
- Do not set the system too conservatively, because this will lead to underutilization of the infrastructure and unnecessarily rapid filling of buffers (keep in mind the queue protection strategies for buffers);
- Test hypotheses that show that the system and the separate components operate correctly for situation-specific cases.

9.3 Recommendations for future research

In practice, there is a need for system optimal control that takes user interests into account. This dissertation presents a relatively simple approach that focuses on smaller-scale system optimality and limits potential inconvenience to the individual road user. On a large scale and in a realistic context, network and traffic situations are frequently dynamic and complex. This means that system optimal control becomes a function of various network, traffic and control variables - particularly if road users' interests have to be safeguarded. Additional research could also focus on:

- Making the methods presented rapidly and generically applicable for a large diversity of network and traffic situations;
- Analyzing traffic conditions and their impact on network performance to enable rapid adjustment of the control methods presented to achieve or approach system optimality;
- The impact of faulty measurements and estimates and inaccurate parameter settings on the effect of a control strategy;
- The application of models combined with actual (including real-time) traffic data to predict conditions and thus arrive at the correct settings for the system (duration of rush hours, turn fractions in buffers, etc.);
- Research into new or improved technologies for providing correct estimates of conditions on the underlying and freeway road network and anticipating traffic problems.

This section presents categories of topics for follow-up research: Control strategies, Analysis for network-wide traffic management, Practical operationalization.

9.3.1 Recommendations for research on control strategies

Follow-up research on the topics below would complement the contributions presented in this dissertation:

Service level-based routing (Chapter 3)

- **Prevention of spill back on a route with multiple bottlenecks.** Research is needed into how to distribute traffic on a route with multiple bottlenecks in such a way as to prevent spill back at the bottleneck with the greatest impact on the network performance;
- **Further development of the method.** Additional research is also needed into the role that the finite-state machine presented, or any other methods, can play in this;
- **Generalization for application in other domains.** How can this method be used in other domains? Service levels are relevant when available capacity is scarce. Access is essential for emergency or other services that provide regional assistance and the strategic spread of demand over the available capacity is relevant from the points of view of both the system and the user.

Coordinated control strategies (Chapters 5 and 6)

- **The practical preconditions that have to be met for algorithms to function well.** Further research is needed to identify the impact of data quality, traffic flow characteristics and infrastructure layouts on the effectiveness and execution of the proposed control strategies. For example:
  - The setup needed for the physical infrastructure, as the size of buffers on on-ramps and intersections has a direct effect on the solvability of a traffic problem and how the control strategy functions;
  - The sensitivity of a control system to faulty measurements and estimates, as faulty measurements and fluctuating traffic demand lower the performance of the algorithm.

9.3.2 Recommendations for analysis for network-wide traffic management

Despite the steps we have made in this thesis in the development of network-wide traffic management, there are several research directions that require further exploration. Let us briefly describe these.

Cumulative curves (Chapter 4)

- **Operationalization for different forms of coordination.** To make the method with cumulative curves practically applicable for random forms of coordination and integration, the approach will have to be further generalized and automated;

- **Application of model predictions and historic data** to analyze the costs and benefits of coordination in advance and to determine optimal buffer sets, including the filling strategy;

- **Real-time application** for determining clusters of buffers that can be filled in series while the control demand (in or outflow reduction) can always be fulfilled.

Systematic tuning of feedback gains (Chapter 6)

- **Generalization of the method.** The method is based on a specific state space model, which is not easy to apply in practice. To this end, an interface will have to be developed which translates a network situation to the state space model used in this study;

- **Validation of the outcomes.** The method should, moreover, be validated in practical situations and more complex model environments in which the state of the traffic is more stochastic and changeable.

9.3.3 Recommendations for research on further practical operationalization

In this final part, we explore briefly the final recommendations regarding research that is needed for practical implementation of the various concepts presented in this dissertation.

Control and estimation conditions in practice (Chapters 7 and 8)

- **Anticipating flow breakdowns.** This is terrain for further research because it is difficult to have the system kick in at exactly the right moment. Unnecessary or premature activation and settings that are too conservative lead to the filling up of the available buffer capacity and travel time losses;

- **Capacity estimation on freeways.** In practice, it is important that freeway capacity is adequately estimated. Additional research is needed to this end because both over and underestimates lead to suboptimal network performance due to capacity drop, underutilization of the freeway and unnecessary buffering of traffic;

- **Queue length estimation detectors on the underlying road network.** Further research needs to be carried out on less expensive techniques for determining queues accurately with minimal loop configurations. The correct estimation of queues determines the effectiveness of the control system, obtaining explainable behavior from the controllers and realizing stable network conditions;

- **Functional aspects of an integrated control system.** More detailed research is also required into key functional aspects such as:

  - The way in which buffered traffic has to be released if buffers are full;
  - How the system can effectively anticipate bottlenecks;
  - How and when the system activates and deactivates;
  - What network-wide strategies are worthwhile testing, based on the average network state identified by the MFD.

Bibliography

Allsop, R. (1971) SIGSET: A computer program for calculating traffic capacity of signal controlled road junctions, *Traffic Engineering and Control*, 12, pp. 58–60.

Allsop, R. (1976) SIGCAP: A computer program for assessing the traffic capacity of signal controlled road junctions, *Traffic Engineering and Control*, 17, pp. 38–41.

ARANE (2009) Dynamic Traffic Management Vision Mid-Netherlands, Tech. rep., ARANE.

Athans, M. (1978) Advances and open problems on the control of large scale systems, in: *Proceedings of the 7th IFAC Conference*, pp. 2371–2382.

Beenker, N., I. Schelling, M. Schoemaker, B. van Engelenburg, H. Kwakernaat, M. Schreuder (2015) Evaluatie Praktijkproef Amsterdam Wegkant (in Dutch), *Presented at: Nationaal Verkeerskundecongres*.

Bellemans, T., B. D. Schutter, B. D. Moor (2006) Model predictive control for ramp metering of motorway traffic: A case study, *Control Engineering Practice*, 14(7), pp. 757–767.

Boillot, F., S. Midenet, J. Pierrelee (2006) The real-time urban traffic control system CRONOS: Algorithm and experiments, *Transportation Research Part C: Emerging Technologies*, 14, pp. 18–38.

Cassidy, M., R. Bertini (1999) Some traffic features at freeway bottlenecks, *Transportation Research Part B: Methodological*, 33, pp. 25–42.

Cassidy, M., J. Rudjanakanoknad (2005) Increasing the capacity of an isolated merge by metering its on-ramp, *Transportation Research Part B: Methodological*, 39(10), pp. 896–913.

Chu, L., H. Liu, W. Recker, H. Zhang (2004) Performance evaluation of adaptive ramp-metering algorithms using microscopic traffic simulation model, *Journal of Transportation Engineering*, 130(3), pp. 330–338.

Daganzo, C., N. Geroliminis (2008) An analytical approximation for the macroscopic fundamental diagram of urban traffic, *Transportation Research Part B: Methodological*, 42(9), pp. 771–781.

de Oliveira, L., E. Camponogara (2010) Multi-agent model predictive control of signaling split in urban traffic networks, *Transportation Research Part C: Emerging Technologies*, 18(1), pp. 120–139.

Diakaki, C., M. Papageorgiou, K. Aboudolas (2002) A multivariable regulator approach to traffic-responsive network-wide signal control, *Control Engineering Practice*, 10(2), pp. 183–195.

Gallivan, S., B. Heydecker (1988) Optimising the control performance of traffic signals at a single junction, *Transportation Research Part B*, 22(5), pp. 357–370.

Gartner, N. (1983) Opac: A demand-responsive strategy for traffic signal control, *Transportation Research Record*, 1(906), pp. 75–81.

Geroliminis, N., C. Daganzo (2008) Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings, *Transportation Research Part B: Methodological*, 42(9), pp. 759–770.

Gonzalez, P., D. Hardesty, G. Hatcher, M. Mercer, M. Waisley (2012) Integrated corridor management: Implementation guide and lessons learned, Tech. Rep. FHWA-JPO-16-280, U.S. Department of Transportation.

Hall, F., K. Agyemang-Duah (1991) Freeway capacity drop and the definition of capacity, *Transportation Research Record*, 1320(1), pp. 91–98.

Han, B., R. Reiss (1994) Coordinating ramp meter operation with upstream intersection traffic signal, *Transportation Research Record*, 1446, pp. 44–47.

Hegyi, A. (2004) *Model predictive control for integrating traffic control measures*, Ph.D. thesis, Delft University of Technology.

Hegyi, A., S. Hoogendoorn (2010) Dynamic speed limit control to resolve shock waves on freeways: Field test results of the SPECIALIST algorithm, in: *Proceedings of the 13th IEEE International Conference on Transportation Systems*.

Hegyi, A., S. Hoogendoorn, M. Schreuder, H. Stoelhorst (2009) The expected effectivity of the dynamic speed limit algorithm SPECIALIST: A field data evaluation method, in: *Proceedings of the European Control Conference*.

Hegyi, A., B. D. Schutter, H. Hellendoorn (2005a) Model predictive control for optimal coordination of ramp metering and variable speed limits, *Transportation Research Part C: Emerging Technologies*, 13(3), pp. 185–209.

Hegyi, A., B. D. Schutter, J. Hellendoorn (2005b) Optimal coordination of variable speed limits to suppress shock waves, *IEEE Transactions on Intelligent Transportation Systems*, 6(1), pp. 102–112.

Henry, J.-J., J. L. Farges, J. Tuffal (1984) The prodyn real time traffic algorithm, in: *Control in Transportation Systems*, Elsevier, pp. 305–310.

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

Hoogendoorn, S., R. Landman, J. van Kooten, M. Schreuder, R. Adams (2015) Design and implementation of an integrated network management methodology in a regional network, *Transportation Research Record*, 2489(1), pp. 20–28.

Hoogendoorn, S., J. van Kooten, R. Adams (2016) Lessons learned from field operational test of integrated network management in Amsterdam, *Transportation Research Record: Journal of the Transportation Research Board*, 12(2554), pp. 111–119.

Horowitz, R., A. May, A. Skabardonis, P. Varaiya, H. Zhang, G. Gomes, L. Munoz, X. Sun, D. Sun (2005) Design, field implementation and evaluation of adaptive ramp metering algorithms, Tech. Rep. UCB-ITS-PRR-2005-2, University of California PATH Program.

Hourdakis, J., P. G. Michalopoulos (2002) Evaluation of ramp control effectiveness in two twin cities freeways, *Transportation Research Record*, 1811(1), pp. 21–29.

Hunt, P., D. Robertson, R. Bretherton (1982) The SCOOT on-line traffic signal optimisation technique, *Traffic Engineering & Control*, 23(4), pp. 190–192.

Jacobson, L., K. Henry, O. Mehyar (1989) Real-time metering algorithm for centralized control, *Transportation Research Record*, 1232(1), pp. 17–26.

Karimi, A., A. Hegyi, B. de Schutter, J. Hellendoorn, F. Middelham (2004) Integrated model predictive control of dynamic route guidance information systems and ramp metering, in: *Proceedings of the 7th IEEE International Conference on Intelligent Transportation Systems*, pp. 491–496.

Knoop, V., S. Hoogendoorn, H. van Zuylen (2007) Quantification of the impact of spillback modeling in assessing network reliability, in: *Transportation Research Board 86th Annual Meeting*, vol. 07-0859.

Kotsialos, A., M. Papageorgiou (2004) Efficiency and equity properties of freeway network-wide ramp metering with amoc, *Transportation Research Part C: Emerging Technologies*, 12(6), pp. 401–420.

Kotsialos, A., M. Papageorgiou, M. Mangeas, H. Haj-Salem (2002) Coordinated and integrated control of motorway networks via non-linear optimal control, *Transportation Research Part C: Emerging Technologies*, 10(1), pp. 65–84.

Kotsialos, A., M. Papageorgiou, A. Messmer (1999) Optimal coordinated and integrated motorway network traffic control, in: *Proceedings of the 14th International Symposium on Transportation and Traffic Theory*, pp. 621–644.

Kotsialos, A., M. Papageorgiou, A. Messmer, F. Middelham, M. Westerman (1998) DACCORD - development and application of co-ordinated control of corridors, Tech. rep., Commission of the European Communities - Advanced Transport Telematics, Brussels, Belgium.

Kotsialos, A., M. Papageorgiou, F. Middelham (2001) Optimal coordinated ramp metering with advanced motorway optimal control, *Transportation Research Record*, 1748(1), pp. 55–65.

Kwon, E., R. Ambadipudi, J. Bieniek (2003) Adaptive coordination of ramp meter and intersection signal for optimal management of freeway corridor, in: *Proceedings of the 82nd Annual Meeting of the Transportation Research Board*.

Kwon, E., S. Nanduri, R. Lau, J. Aswegan (2001) Comparative analysis of operational algorithms for coordinated ramp metering, *Transportation Research Record*, 1748(1), pp. 144–152.

Landman, R., A. Hegyi, S. Hoogendoorn (2015) Coordinated ramp metering based on on-ramp saturation time synchronisation, *Transportation Research Record*, 2484(1), pp. 50–59.

Landman, R., T. Schreiter, A. Hegyi, J. van Lint, S. Hoogendoorn (2012) Policy-based, service level oriented route guidance in road networks, *Transportation Research Record*, 2278(1), pp. 115–124.

Lin, S., B. D. Schutter, Y. Xi, H. Hellendoorn (2012) Efficient network-wide model-based predictive control for urban traffic networks, *Transportation Research Part C: Emerging Technologies*, 24, pp. 122–140.

Lipp, L., L. Corcoran, G. Hickman (1991) Benefits of central control for denver ramp metering system, *Transportation Research Record*, 1320(1), pp. 3–6.

Little, J., M. Kelson, N. Gartner (1981) Maxband: A program for setting signals on arteries and triangular networks, *Transportation Research Record*, 1(795), pp. 40–46.

Mac Carley, C., S. Mattingly, M. Mc Nally, D. Mezger, J. Moore (2002) Field operational test of integrated freeway ramp metering / Arterial adaptive signal control - Lessons learned in Irvine, California, *Transportation Research Record*, 1811(1), pp. 76–83.

Mahajan, N., A. Hegyi, G. van de Weg, S. Hoogendoorn (2015) Integrated variable speed limit and ramp metering control against jam waves: A COSCAL v2 based approach, in: *Proceedings of the 18th IEEE International Conference on Intelliget Transportation Systems*.

Mammar, S., A. Messmer, P. Jensen, M. Papageorgiou, H. Haj-Salem, L. Jensen (1996) Automatic control of variable message signs in aalborg, *Transportation Research Part C: Emerging Technologies*, 4(3), pp. 131–150.

Messmer, A., M. Papageorgiou (1995) Route diversion control in motorway networks via nonlinear optimization, *IEEE Transactions on Control Systems Technology*, 3(1), pp. 144–154.

Messmer, A., M. Papageorgiou, N. Mackenzie (1998) Automatic control of variable message signs in the interurban scottish highway network, *Transportation Research Part C: Emerging Technologies*, 6(3), pp. 173–187.

Middelham, F., H. Taale (2006) Ramp metering in the netherlands: an overview, in: *Proceedings of the 11th IFAC Symposium on Control in Transportation Systems*, vol. 11, p. 267–272.

Middelham, F., H. Taale, B. van Velzen, H. Haj-Salem, M. Papageorgiou, J. Chrisoulakis, P. Gower, D. Tordjman, J. Psarras, T. Mc. Lean (1995) Eurocor experimental results and comparative analysis, Tech. rep., Transport Telematics Office.

Minciardi, R., F. Gaetani (2001) A decentralized optimal control scheme for route guidance in urban road networks, in: *Proceedings of the IEEE Conference on Intelligent Transportation Systems*, pp. 1195–1199.

Muller, H., M. de Leeuw (2006) New method to design traffic control programs, *Transportation Research Record*, 1978(1), pp. 68–75.

of Transport Public Works, D. M. (2013) Beter Geinformeerd op weg: Routekaart 2013 - 2023 - Hoofddocument, Tech. Rep. 483463, Dutch Ministry of Transport Public Works.

of Transport Public Works, D. M., W. Management (2008) Policy framework for utilisation: A pillar of better accessibility (in dutch: Beleidskader benutten), Tech. Rep. PUC130235-31.

Paesani, G., J. Kerr, P. Perovich, F. Khosravi (1997) System wide adaptive ramp metering (SWARM), in: *Proceedings of the 7th Annual Meeting of American ITS*.

Papageorgiou, M. (1990) Dynamic modeling, assignment, and route guidance in traffic networks, *Transportation Research Part B: Methodological*, 24(6), pp. 471 – 495.

Papageorgiou, M. (1995) An integrated control approach for traffic corridors, *Transportation Research Part C: Emerging Technologies*, 3(1), pp. 19–30.

Papageorgiou, M., J. Blosseville, H. Haj-Salem (1990) Modelling and real-time control of traffic flow on the southern part of boulevard périphérique in paris: Part ii: Coordinated on-ramp metering, *Transportation Research Part A: General*, 24(5), pp. 361–370.

Papageorgiou, M., C. Diakaki, V. Dinopoulou, A. Kotsialos, Y. Wang (2003) Review of road traffic control strategies, *Proceedings of the IEEE Conference*, 91(12), pp. 2043–2067.

Papageorgiou, M., H. Haj-Salem, J. Blosseville (1991) Alinea: A local feedback control law for on-ramp metering, *Transportation Research Record*, 1320(1), pp. 58–64.

Papageorgiou, M., A. Kotsialos (2002) Freeway ramp metering: An overview, *IEEE Transactions on Intelligent Transportation Systems*, 3, pp. 271–281.

Papamichail, I., A. Kotsialos, I. Margonis, M. Papageorgiou (2010a) Coordinated ramp metering for freeway networks - a model-predictive hierarchical control approach, *Transportation Research Part C: Emerging Technologies*, 18(3), pp. 311–331.

Papamichail, I., M. Papageorgiou (2008) Traffic-responsive linked ramp-metering control, *IEEE Transactions on Intelligent Transportation Systems*, 9(1), pp. 111–121.

Papamichail, I., M. Papageorgiou, V. Vong, J. Gaffney (2010b) Heuristic ramp-metering coordination strategy implemented at Monash freeway, Australia, *Transportation Research Record*, 2178(1), pp. 10–20.

Pavlis, Y., M. Papageorgiou (1999) Simple decentralized feedback strategies for route guidance in traffic networks, *Transportation Science*, 33(3), pp. 264–278.

Peeta, S., H. Mahmassani (1995) System optimal and user equilibrium time-dependent traffic assignment in congested networks, *Annals of Operations Research*, 60(1), pp. 81–113.

Pooran, F., F. Lieu, C. Henry (1994) Evaluation of system operating strategies for ramp metering and traffic signal coordination, in: *Proceedings of the Annueal meeting of the Intelligent Vehicle Highway Society America*.

Recker, W. (2003) Development of an adaptive corridor traffic control model, Tech. Rep. 109403211, University of California PATH program.

Rijkswaterstaat (2002) Handbook sustainable traffic management, Tech. Rep. ISBN903693625X, AVV Transport Research Centre.

Robertson, D. (1969) Transyt - method for area traffic control, *Traffic Engineering & Control*, 11(6), pp. 276–281.

Savelberg, F., J. Korteweg (2011) Slim benutten: bereikbaarheidsmaatregelen op een rij, Tech. rep., Kennisinstituut voor Mobiliteitsbeleid (KIM).

SBVV (2011) Eind advies, Tech. rep., Strategisch Beraad Verkeersinformatie en Verkeersmanagement.

Schreiter, T., R. Landman, J. van Lint, A. Hegyi, S. Hoogendoorn (2012) Vehicle class-specific route guidance of freeway traffic by model-predictive control, *Transportation Research Record*, 7(2324), pp. 53–62.

Sen, S., L. Head (1997) Controlled optimization of phases at an intersection, *Transportation Science*, 31(1), pp. 5–17.

Silcock, J., A. Sang (1990) SIGSIGN: A phase-based optimisation program for individual signal-controlled junctions, *Traffic Engineering and Control*, 31, pp. 291–98.

Sims, A., K. Dobinson (1980) The Sydney coordinated adaptive traffic (SCAT) system philosophy and benefits, *IEEE Transactions on Vehicular Technology*, 29(2), pp. 130–137.

Smaragdis, E., M. Papageorgiou, E. Kosmatopoulos (2004) A flow-maximizing adaptive local ramp metering strategy, *Transportation Research Part B: Methodological*, 38(3), pp. 251–270.

Spiliopoulou, A., D. Manolis, I. Papamichail, M. Papageorgiou (2010) Queue management techniques for metered freeway on-ramps, *Transportation Research Record*, 2178(1), pp. 40–48.

Stephanedes, Y. (1994) Implementation of On-line Zone Control Strategies for Optimal Ramp Metering in the Minneapolis Ring Road, in: *Proceedings of the 7th International Conference on Road Traffic Monitoring Control*, pp. 181–184.

Stephanedes, Y., K. Chang (1993) Optimal control of freeway corridors, *Journal of Transportation Engineering*, 119, pp. 504–514.

Stevanovic, A. (2010) Adaptive traffic control systems: Domestic and foreign state of practice, Tech. rep., Nationational Cooperative Highway Research Program (NCHRP).

Su, D., X. Lu, R. Horowitz, Z. Wang (2014) Coordinated ramp metering and intersection signal control, *International Journal of Transportation Science and Technology*, 3, pp. 179–192.

Taale, H., W. Schouten, J. van Kooten (1994) Design of a coordinated ramp-metering system near Amsterdam, in: *Proceedings of the 7th International Conference on Road Traffic Monitoring and Control*, pp. 185–189.

Taale, H., M. Westerman (2005) The application of sustainable traffic managment in the Netherlands, in: *Proceedings of the European Transport Conference*.

Tampère, C., R. Corthout, D. Cattrysse, L. Immers (2011) A generic class of first order node models for dynamic macroscopic simulation of traffic flows, *Transportation Research Part B: Methodological*, 45, pp. 289–309.

Tian, Z., K. Balke, R. Engelbrecht, L. Rilett (2002) Integrated control strategies for surface street and freeway systems, *Transportation Research Record*, 1811, pp. 92–99.

Tian, Z., C. Messer, K. Balke, T. Urbanik (2005) Integration of diamond interchange and ramp metering operations, *Transportation Research Record*, 1995, pp. 101–111.

Treiber, M., A. Kesting, E. Wilson (2011) Reconstructing the traffic state by fusion of heterogeneous data, *Computer-Aided Civil and Infrastructure Engineering*, 26(6), pp. 408–419.

Urbanik, T., D. Humphreys, B. Smith, S. Levine (2006) Coordinated freeway and arterial operations handbook, Tech. rep., FHWA.

van Aerde, M., S. Yagar (1988) Dynamic integrated freeway traffic signal networks - problems and proposed solutions, *Transportation Research Part A: Policy and Practice*, 22(6), pp. 435–443.

van de Weg, G., A. Hegyi, H. Hellendoorn, S. Shladover (2014) Cooperative systems based control for integrating ramp metering and variable speed limits, in: *Proceedings of the 93th Annual Meeting of the Transportation Research Board*.

van den Berg, M., B. B. De Schutter, A. Hegyi, 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.

van den Berg, M., A. Hegyi, B. D. Schutter, J. Hellendoorn (2007) Integrated traffic control for mixed urban and freeway networks: A model predictive control approach, *European Journal of Transport and Infrastructure Research*, 7, pp. 223–250.

van der Veen, F., W. Drewes, G. van der Heijden (2010) Benutten rond Amsterdam, evaluatie van het maatregelpakket verbetering doorstroming a10, Tech. rep., DHV B.V. Amersfoort.

van Kooten, J., K. Adams (2011) Handbook Sustainable Traffic Management Plus (in Dutch), Tech. rep., CROW.

van Kooten, J., H. Meurs (2009) Proof of concept field operational test Amsterdam (in Dutch)), Tech. rep., Dutch Ministry of Transport Public Works and Water Management.

van Lint, J. (2010) Empirical evaluation of new robust travel time estimation algorithms, *Transportation Research Record*, 2160(1), pp. 50–59.

van Lint, J., S. Hoogendoorn, M. Schreuder (2008) Fastlane: New multiclass first-order traffic flow model, *Transportation Research Record*, 2088(1), pp. 177–187.

van Lint, J., T. Schreiter, Y. Yuan (2009) Technische en functionele haalbaarheid check-algoritme voor de productie van statistische verkeersgegevens en indicatoren., Tech. rep., Delft University of Technology.

Vincent, R., C. Young (1986) Self-optimizing traffic signal control using microprocessor: The TRRL MOVA strategy for isolated intersections, *Traffic Engineering and Control*, 27, pp. 385–387.

Vito, M., C. D. Taranto (1990) UTOPIA, in: *Control, computers, communications in transportation*, IFAC Symposia Series, Elsevier, pp. 245–252.

Volp, C., J. Leeuwenburgh, A. Hendriksen, R. Sellies, I. van der Hee (2006) Werkboek regelscenario's - voor gebiedsgericht operationeel verkeersmanagement, Tech. rep., CROW.

Wang, Y., M. Papageorgiou, A. Messmer (2001) Feedback and iterative routing strategies for freeway networks, in: *Proceedings of the IEEE International Conference on Control Applications*, pp. 1162–1167.

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

Wang, Y., J. Vrancken, M. dos Santos Soares (2009) Intelligent network traffic control by integrating top-down and bottom-up control, in: *Proceedings of the Chinese Control and Decision Conference*.

Wang, Y., M. Vrancken, M. Vale, M. Davarynejad (2010) Integration of urban and freeway network control by using a scenario coordination module, in: *Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems*.

Wattleworth, J. A. (1965) Peak-period analysis and control of a freeway system, *Highway Research Record*, 157, pp. 1–21.

Yuan, K., V. Knoop, S. Hoogendoorn (2015) Capacity drop: A relation between the speed in congestion and the queue discharge rate, *Transportation Research Record*, 2491, pp. 72–80.

Zegeye, S. K., B. De Schutter, H. Hellendoorn, E. Breunesse (2009) Reduction of travel times and traffic emissions using model predictive control, in: *American Control Conference*, pp. 5392–5397.

Zhang, H., J. Ma, Y. Nie (2009) Local synchronisation control scheme for congested interchange area in freeway corridor, *Transportation Research Record*, 2129, pp. 173–183.

Zhang, L., D. Levinson (2004) Optimal freeway ramp control without origin-destination information, *Transportation Research Part B: Methodological*, 38(10), pp. 869–887.

Zuurbier, F. (2010) *Intelligent Route Guidance*, Ph.D. thesis, Delft University of Technology.

Zuurbier, F., H. van Zuylen, S. Hoogendoorn, Y. Chen (2006) Generating optimal controlled prescriptive route guidance in realistic traffic networks; a generic approach, *Transportation Research Record*, 1944, pp. 58–66.

Summary

Modular framework to operationalize traffic management on a network level

Due to the ever-growing traffic demand, urban areas all over the world are facing serious congestion problems. To mitigate the negative impacts of congestion, for many years more roads were built and traffic management measures locally implemented. Regardless all the extra asphalt and local control solutions, nowadays demands are often still exceeding the network capacity at more and more locations within a network. Which led us to the following situation: by solving one bottleneck, others might be easily activated elsewhere. The current focus of traffic management has therefore been put on realizing collaboration between traffic management measures that deal with traffic problems from a network perspective.

However, in the operational field of traffic management there are more objectives to satisfy than just improving the overall network performance. This is related to the many different stakeholders involved in the process of formulating a vision upon the functional use of a road network, as well as our spirit of the times that emphasizes the value of the livability of our environment. Therefore, when planning for the improvement of network performance, it is increasingly important to take note of the many different stakeholder interests. With the stakeholders agreeing on a common vision, systems are needed that are able to operationalize the vision based on real-time conditions at the involved freeways and urban roads.

The currently available control methods can rarely be deployed effectively and efficiently on a large scale (e.g. optimizing control strategies and integrated control scenarios), because of their characteristic vulnerabilities.

Their theoretical and operational complexity make these approaches error-prone, expensive to maintain, difficult to operate and hard to apply in real-time.

This dissertation thesis aims at designing new traffic control strategies and algorithms that can be combined into a framework such that they solve problems from a network perspective, integrating traffic measures, policy objectives, user interests and practical applicability. A framework has been developed that efficiently operationalizes integrated network management, along with an elaboration on effective control strategies, associated algorithms and overall design approach. Moreover, the framework shows how control and monitoring solutions can be integrated and tailored to the specific needs of any regional network. It also allows new monitoring and control units to be easily adopted and to operationalize the system stage-wise in practice.

In order to define the research needs more clearly, the focus within the literature survey was put on the deployability of existing control approaches. In other words, we looked at their ability to realize target states in line with policy objectives and as such their network performance enhancing qualities. Based on these findings, preferred controller types (i.e. feedback, feed forward, optimal control) that are suitable for application in an operational context were selected. These existing control approaches were then evaluated to find opportunities for further improvement.

Furthermore, the focus was put on the development of effective heuristic control strategies that are based on a thorough understanding of the root causes of, and the solutions for, undesired traffic phenomena in the network. Although optimal control-based approaches are superior in terms of flexibility and robustness, for practical applications the computational complexities and lack of transparency are sufficiently serious disadvantages to consider heuristic but still generic approaches. In this research optimal control has nonetheless been used to benchmark with system optimality.

Since traffic networks consist of both freeway and urban roads, coordinated or integrated control solutions have been designed for both types of networks. This includes the design of an approach to systematically tune and configure the feedback gains of the used feedback control laws. More-

over, a framework has been presented that allows for integrated application of different control and monitoring units for network wide traffic management.

Finally, designing an integrated network management system includes the use of traffic simulation models for the purpose of understanding, testing, improving and validating the overall control system and its individual components. In order to ensure their well functioning a structured and feedback-based design flow has been defined that keeps the conceptual, technical and functional tests on the system comprehensible and manageable.

In practice, there is a need for an integrated, network-wide control system that can combine state-of-the-art measurement and control methods to identify traffic problems promptly and tackle its root causes. This research has proposed various comprehensible and practical improvements when coordinating traffic management measures such as route guidance, intersection control and ramp metering. The main conclusions are as follows:

- If traffic is routed based on predefined service levels, route capacity within the network can be fully utilized and phenomena that decrease the network performance prevented. Because the controllers observe maximal travel time differences between the routes, excessive delays for rerouted road users remain limited.

- It is shown that the network performance is impacted by the selection of coordinated buffers and the order in which they are filled. For coordinated ramp metering it is beneficial to fill ramps in downstream order and for coordinated intersection control the filling strategy depends on the fraction of traffic towards the bottleneck. Intersection buffers should be filled simultaneously when these fractions are equal and sequentially when fractions differ significantly (starting with the buffers having the largest fraction).

- It is vital that the set of buffers for coordination are chosen carefully. The number and deployment capacity of the buffers required to achieve maximal improvement of the network performance are sit-

uation specific. Selection depends on factors such as the duration and degree of overload of the bottleneck, the size of the freeway capacity drop and the fraction of traffic in the buffer towards the bottleneck.

The previously mentioned filling strategies can be operationalized by means of feedback controllers, which makes them a very suitable and effective solution within the context of network management. When the involved feedback gains are systematically tuned, their impact is maximized and a lot of time for setting them up saved.

To conclude, policy objectives with respect to improving network performance, equity and livability can be put into practice, if they are translated into clear target states that can be maintained by the control system (i.e. maximum travel time differences, flow and density targets or buffer capacity). Overall we can state that coordinated control strategies and the associated algorithms have to meet certain preconditions to be applicable in the existing operational practice. They need to be comprehensible, scalable and have the ability to cope with unforeseen changes in traffic demand and network supply.

R.L. Landman

Samenvatting

Modulair raamwerk om verkeersmanagement op netwerkniveau te operationaliseren

Door de alsmaar groeiende verkeersvraag hebben stedelijke gebieden over de hele wereld te kampen met ernstige congestieproblemen. Om de negatieve gevolgen van congestie te mitigeren, zijn jarenlang meer wegen aangelegd en maatregelen voor verkeersmanagement lokaal geïmplementeerd. Ondanks alle extra asfalt en lokale beheersoplossingen overstijgt de huidige verkeersvraag vaak nog steeds de netwerkcapaciteit op steeds meer plaatsen binnen een netwerk. Dit leidde tot de volgende situatie: door een knelpunt op te lossen, kunnen andere knelpunten gemakkelijk elders worden geactiveerd. De huidige focus van verkeersmanagement is daarom gelegd op het realiseren van samenwerking tussen maatregelen die verkeersproblemen vanuit een netwerkperspectief aanpakken.

Bij operationeel verkeersmanagement zijn er echter meer doelstellingen waaraan moet worden voldaan dan alleen het verbeteren van de netwerkprestatie. Dit heeft enerzijds te maken met het feit dat er veel belanghebbende partijen zijn betrokken bij het formuleren van een visie op het functionele gebruik van een wegennet, en anderzijds met onze tijdgeest die meer de nadruk legt op de waarde van een leefbare omgeving. Kortom, naast het verbeteren van de netwerkprestatie wordt het steeds belangrijker rekening te houden met andere beleidsdoelstellingen. Uiteraard zijn er vervolgens systemen nodig die in staat zijn om de visie te operationaliseren op basis van real-time toestanden op de betrokken snelwegen en stadswegen.

De momenteel beschikbare regelmethoden kunnen lastig effectief en efficient op grote schaal worden toegepast (bv. optimaliserende regelstrategieën en integrerende regelscenario's), vanwege hun karakteristieke kwetsbaarheden. Theoretische en operationele complexiteit maken deze aanpakken foutgevoelig, duur in onderhoud, moeilijk te bedienen en moeilijk in real-time toe te passen.

Tijdens dit onderzoek zijn we opzoek gegaan naar nieuwe regelmethoden voor verkeersmanagement waarbij aandacht is voor aspecten als het oplossen van problemen vanuit een netwerkperspectief, het integreren van verkeersmaatregelen, beleidsdoelstellingen, gebruikersbelangen en praktische toepasbaarheid. Er is een raamwerk ontwikkeld dat integraal verkeersmanagement doeltreffend operationaliseert in de praktijk, inclusief een uitwerking van effectieve meet- en regelstrategieën met bijbehorende algoritmiek en globale ontwerpaanpak. Zo kan het systeem worden afgestemd op de specifieke behoeften van een regionaal verkeersnetwerk, waarbij het stapsgewijs operationeel gemaakt kan worden.

Om de onderzoeksbehoeften te specificeren, is de nadruk tijdens het literatuuronderzoek gelegd op de toepasbaarheid van bestaande regeltechnieken in de praktijk. Met andere woorden, we hebben hun vermogen beschouwd om gewenste toestanden in het netwerk te creëren die in lijn liggen met beleidsdoelstellingen als mede hun vermogen de netwerkprestatie te verbeteren. Op basis van deze bevindingen zijn verschillende regeltechnieken (v.b. feedback, feedforward, optimal control) geselecteerd die in een operationele context toegepast kunnen worden. Deze bestaande regeltechnieken zijn vervolgens geëvalueerd om mogelijkheden voor verdere verbetering te identificeren.

Er gezocht naar doeltreffende heuristische regelstrategieën die gebaseerd zijn op een grondig begrip van de diepere oorzaken van en de oplossingen voor ongewenste verkeersverschijnselen in het netwerk. Hoewel optimale regelaars superieur zijn in termen van flexibiliteit en robuustheid, zijn voor praktische toepassingen de rekenkundige complexiteit en het gebrek aan doorzichtigheid voldoende reden om heuristische maar nog steeds generieke benaderingen te overwegen. Optimaliserende regelingen zijn desalniettemin

toegepast om de prestatie van de voorgestelde aanpakken te vergelijken met systeemoptimaliteit.

Aangezien verkeersnetwerken zowel uit snelwegen als uit stadswegen bestaan, zijn voor beide typen netwerken gecoördineerde en geïntegreerde regelaars ontworpen, inclusief een aanpak om op systematische wijze de feedback gains te kalibreren van voorgestelde feedback regelaars. Bovendien is een raamwerk gepresenteerd dat de geïntegreerde toepassing van verschillende meet- en regeleenheden voor netwerkbreed verkeersbeheer mogelijk maakt.

Het ontwerpen van een geïntegreerd netwerkbreed regelsysteem vereist het gebruik van simulatiemodellen om het totale regelsysteem en de afzonderlijke componenten te leren begrijpen, te testen en verbeteren en te valideren. Om een juiste werking ervan te garanderen, is een gestructureerd en op feedback gebaseerd ontwerpproces gedefinieerd dat het conceptuele, technische en functionele testen van het systeem begrijpelijk en beheersbaar houdt.

In de praktijk is er behoefte aan een netwerkbreed regelsysteem dat de nieuwste meet- en regeltechnieken kan combineren om verkeersproblemen snel op te sporen en de onderliggende oorzaken aan te pakken. In dit onderzoek zijn verschillende begrijpelijke en praktische verbeteringen voorgesteld voor het coördineren van maatregelen zoals routegeleiding, verkeerslichtregelingen en toeritdosering. De belangrijkste conclusies luiden als volgt:

- Als het verkeer wordt geleid op basis van vooraf gedefinieerde serviceniveaus, kan de routecapaciteit binnen het netwerk volledig worden benut en kunnen verschijnselen die de netwerkprestaties verminderen worden voorkomen. Omdat de regelaars een maximaal reistijdverschil bewaken tussen routes, blijven buitensporige vertragingen voor omgeleide weggebruikers beperkt.
- Aangetoond wordt dat de netwerkprestatie wordt beïnvloed door de selectie van gecoördineerde buffers en de volgorde waarin deze worden gevuld. Voor gecoördineerde toeritdosering is het gunstig om de toeritten in stroomafwaartse volgorde te vullen en voor gecoördineerde verkeerslichtregelingen hangt de vulstrategie van buffers af van de fractie van het verkeer dat zich vanuit de buffer richting het knelpunt

beweegt. Kruispuntbuffers dienen gelijktijdig gevuld te worden als deze fracties gelijk zijn, en achtereenvolgens als de fracties aanzienlijk verschillen (te beginnen met de buffers met de grootste fractie).

- Het is verder van belang dat de buffers voor coördinatie zorgvuldig worden gekozen. Het optimale aantal en hun opstelcapaciteit zijn situatie-specifiek bij het nastreven van verbetering van de netwerkprestatie. Selectie hangt af van factoren zoals de duur en de mate van overbelasting van het knelpunt, de grootte van de capaciteitsval op de snelweg en de fractie van het verkeer in de buffer naar het knelpunt toe.

De eerder genoemde vulstrategieën kunnen worden geoperationaliseerd met behulp van feedbackregelaars. Dit soort regelaars zijn dus een zeer geschikte en effectieve oplossing in de context van verkeersmanagement. Wanneer de betrokken feedback gains systematisch worden bepaald, dan wordt het effect van de regelaar gemaximaliseerd en veel tijd bespaard voor het instellen.

Tot slot kunnen beleidsdoelstellingen met betrekking tot het verbeteren van netwerkprestaties, gelijkheid en leefbaarheid in de praktijk worden gebracht, als ze worden vertaald in duidelijke toestanden die door het regelsysteem kunnen worden gehandhaafd (d.w.z. maximale reistijdverschillen, gewenste toestanden met betrekking tot doorstroming en dichtheid, of buffercapaciteit). In het algemeen kunnen we stellen dat gecoördineerde regelstrategieën en de bijbehorende algoritmen aan bepaalde voorwaarden moeten voldoen om in de bestaande operationele praktijk te kunnen worden toegepast. Ze moeten begrijpelijk en schaalbaar zijn en het vermogen hebben om te gaan met onvoorziene veranderingen in de verkeersvraag en het netwerkaanbod.

R.L. Landman

About the author

Ramon Landman (1981) is and always has been fascinated by the nature of logic. He has spent both his childhood, young adolescence and his mature life trying to fathom it. In this quest he has embraced systems approaches and creating synergy between subjects and people. After University, Ramon started off as a scientific consultant and worked on an integral approach for efficient incident management on the Dutch roads to minimize response times of the aid services. In this trajectory he was especially interested in finding ways to create maximum impact by integrating both the knowledge and interests of involved stakeholders. In his next step, as a researcher and during his PhD he studied coordinated control solutions to manage large-scale road networks from a network perspective. He found himself captured by the necessity of using comprehensible concepts and strategies that could be combined into a framework that is able to control the network integrally with finesse. He then went on discovering a whole new area related to logic and synergy: namely systems of people. This is when he enrolled in the field of systems coaching and collective leadership. He was taught body based methods to identify people their talents, and how to facilitate people to step forward or step backwards in the right time to maximizing a team's performance. The latter led him to fulfil the position of director for a small-scale engineering company that specializes in the development of talent. Within this role he combines his engineering skills and knowledge with an eye for the person behind the technician. As shown, his holistic and systems point of view runs like a thread through his life and can certainly be recognised in this dissertation.

Publications

Journal papers

1. Landman, R., T. Schreiter, A. Hegyi, J.W.C. van Lint and S. Hoogendoorn, Policy-based Service Level-oriented, Route Guidance in Road Networks: a Comparison with System and User Optimal Route Guidance, *Transportation Research Record*, Vol. 2278, pp. 115-124, 2012

2. Landman, R., A. Hegyi, S. Hoogendoorn, Service Level-oriented Route Guidance for Overlapping Routes in Road Networks: A Comparison with MPC, In *Proceedings of the 2012 American Control Conference*, pp. 5775-5782, 2012

3. Landman, R., A. Hegyi, S. Hoogendoorn, Service Level-oriented Route Guidance in Road Traffic Networks, In *Proceedings of the 14th IEEE Conference on Intelligent Transportation systems*, pp. 1120-1125, 2011

4. Landman, R., A. Hegyi, S. Hoogendoorn, Urban Storage Space Selection Method for Integrated Control on a Freeway Bottleneck, *Transportation Research Record*, Vol. 2554, pp. 77-88, 2016

5. Landman, R., A. Hegyi, S. Hoogendoorn, On-ramp Selection Methodology for Coordinated Ramp Metering Schemes, In *Proceedings of the 2015 IEEE Conference on Intelligent Transportation Systems and Control*, pp. 1129-1136, 2015.

6. Landman, R., A. Hegyi, S. Hoogendoorn, Coordinated Ramp Metering based on On-ramp Saturation Time Synchronization, *Transportation Research Record*, Vol. 2484, pp. 50-59, 2015

7. Hoogendoorn, S., R. Landman, J. van Kooten, M. Schreuder, R. Adams, Design and Implementation of Integrated Network Management Methodology in a Regional Network, *Transportation Research Record*, Vol. 2489, pp. 20-28, 2015

8. Hoogendoorn, S., R. Landman, J. van Kooten, M. Schreuder, R. Adams, Design and Implementation of Integrated Network Management Methodology in a Regional Network, *Transportation Research Record*, Vol. 2489, pp. 20-28, 2015

9. Hoogendoorn, S., R. Landman, J. van Kooten, M. Schreuder, Integrated Network Management Amsterdam: Control approach and Test Results, In *Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems*, pp. 474-479, 2013

10. Schreiter, T., R. Landman, J.W.C. van Lint, A. Hegyi, S. Hoogendoorn, Vehicle Class-Specific Route Guidance of Freeway Traffic by Model-Predictive Control, *Transportation Research Record*, Vol. 2324, pp. 53-62, 2012

11. Vreeswijk, J.D., R. Landman, E.C. van Berkum, A. Hegyi, S. Hoogendoorn, B. van Arem, Improving the Road Network Performance with Dynamic Route Guidance by Considering the Indifference Band of Road Users, *IET Intelligent Transport Systems*, Vol. 9, Issue 10, pp. 897-906, 2015

12. Lint, H.W.C., R. Landman, Y. Yuan, C. van Hinsbergen, S. Hoogendoorn, Traffic Monitoring for Coordinated Traffic Management - Experiences from the Field Trial Integrated Traffic Management in Amsterdam, In *Proceedings of the 17th IEEE Conference on Intelligent Transportation systems*, pp. 477-482, 2014

TRAIL Thesis Series

The following list contains the most recent dissertations in the TRAIL Thesis Series. For a complete overview of more than 275 titles see the TRAIL website: www.rsTRAIL.nl.

The TRAIL Thesis Series is a series of the Netherlands TRAIL Research School on transport, infrastructure and logistics.

Landman, R.L., *Operational Control Solutions for Traffic Management on a Network Level*, T2021/17, June 2021, TRAIL Thesis Series, the Netherlands

Zomer, L.-B., *Unravelling Urban Wayfinding: Studies on the development of spatial knowledge, activity patterns, and route dynamics of cyclists*, T2021/16, May 2021, TRAIL Thesis Series, the Netherlands

Núñez Velasco, J.P., *Should I Stop or Should I Cross? Interactions between vulnerable road users and automated vehicles*, T2021/15, May 2021, TRAIL Thesis Series, the Netherlands

Duivenvoorden, K., *Speed Up to Safe Interactions: The effects of intersection design and road users' behaviour on the interaction between cyclists and car drivers*, T2021/14, April 2021, TRAIL Thesis Series, the Netherlands

Nagalur Subraveti, H.H.S., *Lane-Specific Traffic Flow Control*, T2021/13, March 2021, TRAIL Thesis Series, the Netherlands

Beirigo, B.A., *Dynamic Fleet Management for Autonomous Vehicles: Learning- and optimization-based strategies*, T2021/12, March 2021, TRAIL Thesis Series, the Netherlands

Zhang, B., *Taking Back the Wheel: Transition of control from automated cars and trucks to manual driving*, T2021/11, February 2021, TRAIL Thesis Series, the Netherlands

Boelhouwer, A., *Exploring, Developing and Evaluating In-Car HMI to Support Appropriate use of Automated Cars*, T2021/10, January 2021, TRAIL Thesis Series, the Netherlands

Li, X., *Development of an Integrity Analytical Model to Predict the Wet Collapse Pressure of Flexible Risers*, T2021/9, February 2021, TRAIL Thesis Series, the Netherlands

Li, Z., *Surface Crack Growth in Metallic Pipes Reinforced with Composite Repair System*, T2021/8, January 2021, TRAIL Thesis Series, the Netherlands

Gavriilidou, A., *Cyclists in Motion: From data collection to behavioural models*, T2021/7, February 2021, TRAIL Thesis Series, the Netherlands

Methorst, R., *Exploring the Pedestrians Realm: An overview of insights needed for developing a generative system approach to walkability*, T2021/6, February 2021, TRAIL Thesis Series, the Netherlands

Walker, F., *To Trust or Not to Trust? Assessment and calibration of driver trust in automated vehicles*, T2021/5, February 2021, TRAIL Thesis Series, the Netherlands

Schneider, F., *Spatial Activity-travel Patterns of Cyclists*, T2021/4, February 2021, TRAIL Thesis Series, the Netherlands

Madadi, B., *Design and Optimization of Road Networks for Automated Vehicles*, T2021/3, January 2021, TRAIL Thesis Series, the Netherlands

Krabbenborg, L.D.M., *Tradable Credits for Congestion Management: support/reject?*, T2021/2, January 2021, TRAIL Thesis Series, the Netherlands

Castelein, B., *Accommodating Cold Logistics Chains in Seaport Clusters: The development of the reefer container market and its implications for logistics and policy*, T2021/1, January 2021, TRAIL Thesis Series, the Netherlands

Huang, B., *The Influence of Positive Interventions on Cycling*, T2020/20, December 2020, TRAIL Thesis Series, the Netherlands

Xiao, L., *Cooperative Adaptive Cruise Control Vehicles on Highways: Modelling and Traffic Flow Characteristics*, T2020/19, December 2020, TRAIL Thesis Series, the Netherlands