Real-Time Operations Planning and Control of High-Frequency Transit

by

Gabriel Eduardo Sánchez-Martínez

M.S. Transportation
Massachusetts Institute of Technology, 2013

B.S. Civil Engineering
University of Puerto Rico, Mayagüez, 2010

Submitted to the Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Transportation

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

Author . . . .
Department of Civil and Environmental Engineering
February 2, 2015

Certified by . . . .
Nigel H.M. Wilson
Professor of Civil and Environmental Engineering
Thesis Supervisor

Certified by . . . .
Haris N. Koutsopoulos
Professor of Civil and Environmental Engineering
Northeastern University
Thesis Supervisor

Accepted by . . . .
Heidi M. Nepf
Donald and Martha Harleman Professor of Civil and Environmental Engineering
Chair, Departmental Committee for Graduate Students

Real-Time Operations Planning and Control of High-Frequency Transit

by

Gabriel Eduardo Sánchez-Martínez

Submitted to the Department of Civil and Environmental Engineering
on February 2, 2015, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Transportation

Abstract

High-frequency transit systems are essential for the socioeconomic and environmental well-being of large and dense cities. The planning and control of their operations are important determinants of service quality. Transit operators are increasingly adopting data collection devices that enable real-time monitoring of vehicle locations and demand, but existing models and current practice limit the utility of this information. This research develops new concepts, frameworks, and models for real-time optimization of operations, utilizing both historical and real-time information originating from connected data collection devices, including automated vehicle location, automated fare collection, and automatic passenger counting systems.

Previous control strategies either do not forecast system states or rely on forecasts based on running times and demand assumed to be static. This research develops an optimization model for holding-based control that incorporates dynamics, producing a holding policy that accounts not only for the current state of the system, but also for expected changes in running times and demand, due to both exogenous and endogenous dynamics. This information advantage can lead to improved performance when a transit service faces typical changes in running times and demand over time, as well as potentially disruptive events such as signal failures, disabled rolling stock, and demand surges. Anticipatory control policies allow the transit service to react before disruptions develop. It is shown that information about dynamics is particularly valuable when it leads to better predictions of capacity being reached.

Although headway and optimization-based control strategies generally outperform schedule-adherence strategies, high-frequency operations are mostly planned with schedules, in part because operators must observe resource constraints (neglected by most control strategies) while planning and delivering service. This research develops a schedule-free paradigm for high-frequency transit operations, in which trip sequences and departure times are optimized in real-time, employing stop-skipping strategies and utilizing real-time information to maximize service quality while satisfying operator resource constraints. Following a discussion of possible methodological approaches, a simple methodology is applied to operate a simulated transit service without schedules. Results demonstrate the feasibility of the new paradigm and suggest possible methodology improvements.

Thesis Supervisor: Nigel H.M. Wilson
Title: Professor of Civil and Environmental Engineering

Thesis Supervisor: Haris N. Koutsopoulos
Title: Professor of Civil and Environmental Engineering, Northeastern University

To my family

Acknowledgments

This research was funded in part by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1122374 and in part by Transport for London. This research also benefited from the support of the Bus Rapid Transit Centre of Excellence, funded by the Volvo Research and Educational Foundations (VREF), and MIT's MISTI-Chile program.

I thank my advisors, professors Nigel Wilson and Haris Koutsopoulos, for their wisdom and insight, for engaging discussions on this and future research, for their continual support and guidance, and for their help editing this dissertation. I thank Professor González Quevedo for encouraging me to consider graduate studies in transportation, and for the invaluable knowledge and skills he has taught me. Professors Jinhua Zhao, Juan Carlos Muñoz, and Carolina Osorio gave suggestions during and after committee meetings that helped shape this research. Fred Salvucci and John Attanucci also contributed. George Kocur taught me how to approach analysis and design of algorithms, skills I found essential for this research.

John Barry from London Buses provided highly valued insight and support. The transportation engineering faculty and staff at Pontificia Universidad Católica de Chile, in particular Juan Carlos Muñoz, Felipe Delgado, Ricardo Giesen, Juan Enrique Coeymans, and Ignacia Torres, received me with warmth and generosity during my one-week visit to Santiago. Discussions with Ricardo Giesen, Felipe Delgado, and Juan Carlos Muñoz during that visit led to the topic of holding with information about dynamics presented in Chapter 2. Ginny Siggia helped with many administrative matters. Daniel Hastings facilitated my stay in MIT housing during the last month of this research, allowing me to focus on finishing this dissertation.

I thank the MITES program for introducing me to the academic intensity of MIT even before starting my undergraduate studies, and the Tech Catholic Community for being so friendly and welcoming. My friends Katie Pincus, Jamie Rosen, Aditi Chandra, Candace Brakewood, Mickaël Schil, Naomi Stein, Laura Viña, Carlos Gutiérrez de Quevedo Aguerrebere, Fabiola Michel, Jay Gordon, Janet Choi, Varun Pattabhiraman, Alex Malikova, Kevin Muhs, Kari Hernández, Caroline Ducas, Harsha Ravichandran, Nina Panagioutidou, Ina Kundu, David Block-Schachter, Cecilia Viggiano, Laura Riegel, Anson Stewart, Farah Machlab, Stephen Tuttle, Yiwen Zhu, Michel Babany, Raphael Dumas, Emily Gates, Lauren Tarte, Anne Halvorsen, Gabriel Goulet-Langlois, David Maltzan, Dan Wood, William Chow, Dominick Tribone, Tatiana Peralta, Anna Matías, Meisy Ortega, Rebecca Heywood, Ofir Hilvert, Alyona Michel, Evelien van der Hurk, Sandy Tenorio, Luis Somoza, Mónica Crespo, Amanda Gaudreau, León Valdés Saavedra, Jorge Elizondo Martínez, Andrea Ríos, and many others made my time at MIT enjoyable with good conversation, a meal shared, or a concert outing. Riley kept me friendly company on many occasions.

I am grateful to my cousins in London, especially Sara and Andrey Kidel, for welcoming me and allowing me to stay with them during my summer in London, to my cousin Rosa Irene Martínez García, who came to my doctoral defense from Providence, and to all my family, especially my parents, Eduardo Sánchez Martínez and María Martínez, and my brother, Javier Sánchez Martínez, for their unending love, encouragement, and support.

Contents

1 Introduction **11**

1.1 Cities and High-Frequency Public Transit 11

1.2 Real-Time Data and Information 12

1.3 Model-Based Operations Control 14

1.4 Research Objectives 19

1.5 Dissertation Structure 19

2 Holding Control with Dynamics **23**

2.1 Literature Review 24

2.2 Framework and Formulation 28

2.2.1 Assumptions 28

2.2.2 Holding Time Optimization 30

2.2.3 Performance Model 31

2.2.4 Cost Model 34

2.2.5 Control Strategy 36

2.2.6 Dimensionality Reduction 36

2.3 Application 37

2.4 Results and Discussion 39

2.4.1 Passenger Cost 40

2.4.2 Headways 42

2.4.3 Crowding at Stops 44

2.4.4 Sensitivity to Stochasticity 44

2.4.5 Sensitivity to Current State Information Accuracy . . . 46

2.4.6 Computation Time 47

2.5 Concluding Remarks 47

3 Holding Control Under Event Driven Dynamics **51**

3.1 Literature Review 52

3.2 Framework 53

3.3 Control Model 55

3.4 Applications 55

3.4.1 Unforeseen Event 56

3.4.2 Foreseen Event 58

3.4.3 Effect of Erroneous Information 62

3.5 Concluding Remarks 63

4 Schedule-Free Operations 65
4.1 Motivation 67
4.2 Literature Review 68
4.3 Framework 70
4.4 Implications 78
4.4.1 Reliance on Information Technologies 78
4.4.2 Methods and Practices 79
4.4.3 Human Factors 81
4.4.4 Operations with Uncertain Resources 83
4.4.5 Operations with Autonomous Fleets 83
4.5 General Methodology 84
4.5.1 Problem Complexity 87
4.6 Simplified Methodology 91
4.6.1 Event-Based Performance Model 96
4.6.2 Exogenous Trip Running Time Model 97
4.6.3 Trip Sequence Optimization 98
4.6.4 Departure Time Optimization 101
4.7 Application 103
4.7.1 Transit Line 105
4.7.2 Results 109
4.7.3 Evaluation of Feasible Sequence Set Cardinality Reduction Heuristics 119
4.8 Concluding Remarks 121

5 Conclusion 125
5.1 Summary 125
5.2 Contributions 130
5.3 Future Work 131

Bibliography 135

Chapter 1

Introduction

High-frequency public transportation systems, those with high enough service frequencies to allow passengers to turn up at their origin stop expecting a short wait rather than having planned to take a specific vehicle trip, are essential for the well-being of large and dense cities. They include bus and rail lines, and serve millions of trips daily (Transport for London, 2014 and MTA New York City Transit, 2014). The planning and control of their operations are important determinants of service quality and of how successfully they accomplish their objectives. This research sets out to develop new concepts, frameworks, and models to improve the effectiveness of operations planning and control by utilizing real-time information originating from connected automated data collection systems.

1.1 Cities and High-Frequency Public Transit

The world's population reached 7.2 billion in 2014 and is expected to increase by a further 2 billion by 2050. More than half of the world's population lives in urban areas, and with rural population expected to decline slightly, the urban population is expected to rise from about 3.9 billion in 2014 to 6.3 billion by 2050. Large cities are growing in number and size. (United Nations, 2014) High densities can increase people's access to diverse job markets, education, and services, but can only be sustained by efficient transportation networks. Congested networks lead to longer and more uncertain trip times, in turn leading to decreased productivity and accessibility. This dampens further growth and encourages people and businesses to relocate to lower density areas, decreasing potential agglomeration benefits. (Graham, 2007b, Graham, 2007a, and Hymel, 2009)

Dense urban areas give rise to spatio-temporally concentrated transportation demand, between specific origins and destinations and along linear corridors. Public transportation systems are most efficient at meeting this demand in terms of cost, space, and energy. Private vehicles require much more space per capita than public transit vehicles, not only while in motion, but also at the destination of trips, where they are often parked for the duration of the trip-generating activity without any direct productivity. The space efficiency of public transport frees up valuable real estate for non-transportation utilization, and frees up road space for trips that cannot be adequately served by public transportation, e.g. freight, which are of great economic importance. It is estimated that over 4 billion gallons of gasoline (or equivalent) are saved and 37 million metric tons of carbon dioxide emissions are avoided annually in the United States due to public transportation, accounting for reduced private passenger vehicle miles, reduced congestion, and reduced travel distances (Neff and Dickens, 2013).

People and businesses make long-term location decisions based on accessibility. While high densities create demand for, and encourage investment in, high-frequency transit, the supply of high-frequency transit foments further densification, thus creating a positive feedback cycle. Dense cities depend on the supply of high-frequency public transit for their socioeconomic and environmental sustainability. For instance, cities like New York and London could not sustain their current densities without their transit networks.

Given the benefits of, and dependence on, high-frequency public transportation systems, the quality of transit operations has significant impact on the functioning of cities. Unreliable operations result in delays and high crowding, diminishing for people and businesses the benefits of being located in a dense city. Many factors affect service quality, including external ones such as traffic (for transit services in shared rights of way) and weather, and internal ones such as unreliability of infrastructure and rolling stock, inadequate resource allocation, and the quality of planning and controlling operations (Sánchez-Martínez, 2012). Large and dense cities such as New York and London are the ones with the longest history of high-frequency public transportation, and therefore the ones with the highest dependence on transit and susceptibility to disruptions due to aging infrastructure.

1.2 Real-Time Data and Information

In parallel with the trend of growing cities, which will make high-frequency transit even more important, information and communications technologies are advancing rapidly and are being increasingly adopted by many sectors, including public transportation. Of particular relevance is the so-called *Internet of things*, describing the growing number of embedded devices connected to the Internet, enabling them to interact with each other, as well as with services and people (Mukhopadhyay and Suryadevara, 2014).

Automated data collection systems have for many years helped transit service providers measure the performance of their services, but connecting their sensors to servers (via the Internet or private networks) enables new real-time applications. Automated vehicle location (AVL) systems, which monitor the locations of vehicles, were originally adopted to improve safety and incident response capability, but are increasingly used for real-time monitoring, control, and passenger information provision. Automated fare collection (AFC) and automatic passenger counting (APC) systems were developed to relieve drivers from the duty of collecting fares and for offline ridership analysis, but also have the potential to provide real-time data useful for load and demand monitoring.

As data is generated, it can be processed to generate useful information. Information can be archived for analysis of trends over time. For example, systematic changes in running times and demand over days of the week and seasons can be quantified along with their stochasticity. Models can combine historical and real-time information, both to estimate the current state of the system and to predict future states. The kinds of data available vary by system.

Vehicle locations can be obtained directly from a real-time stream of AVL data. AVL technology varies by mode and system. Bus AVL systems tend to be based on GPS and odometer readings. Train AVL systems are often based on track circuit occupancy or radio-frequency identification (RFID). Regardless of sensor technology, vehicle positions are now often available in real-time. Depending on the update frequency, and the time elapsed since the latest data for a given vehicle, the current position can be inferred using a vehicle movement model.

Vehicle loads can be estimated with APC data. Vehicle-mounted APC systems count boardings and alightings, from which loads are determined. Rail services often lack in-vehicle APC, but train weighing systems, which are used for braking control, can be used to estimate the load of a train. This can be combined with station passenger counters, such as gateline counts, although this can be an involved process due to the variability of walking time from the gate to the platform and the multiplicity of platforms that can be reached after entering some stations.

Station or stop crowding can be estimated from historical or real-time AFC and APC data. For bus service, where there is no sensing of passengers arriving at stops, historical data can be used to obtain a typical arrival rate (for the season, day of week, and time of day), and a passenger arrival model can be used to estimate the number of passengers waiting to board. The passenger arrival process might simply be assumed to be Poisson, but if the stop is also a transfer station, the model may also use real-time vehicle location data of other routes and historical transfer data to infer arrivals of groups of transferring passengers. For rail stations where passengers interact with gates at entry, station crowding can be estimated by adding a walking time from the gate to the platform. It is necessary to consider vehicle capacity constraints to estimate the number of passengers left at a stop (or platform), because the assumption that the stop or train platform is emptied with each vehicle arrival does not hold when vehicle capacities are binding. If the operator closes gates as part of a metering strategy, it may be useful to have passenger counting before the gateline, perhaps using a video feed.

Passenger destinations can be estimated from historical origin-destination matrices $\mathbf{\Lambda}(t)$ with elements $\lambda_{ij}(t)$ denoting arrival rates at time t of passengers at origin i on their way to destination j. A more sophisticated method involves checking the habitual travel patterns of each passenger observed boarding a vehicle or entering a station. If such a pattern is found, then it may be assumed (with some uncertainty) that the passenger is on his usual trip from i to j. Disaggregate inference methods can be used to generate an origin-destination matrix, or model travel patterns of individuals, if historical fare transaction records are available (Gordon et al., 2013).

If avoiding driver lateness is part of the control objective, duty lateness can be calculated from the current time and the scheduled duty end time, giving how much time is left in the duty. Optionally, spare drivers can be included in the model.

Future arrivals of passengers at a station platform or bus stop, and the destinations of those passengers, can be estimated from historical AFC and APC data, as described previously. Information on future boardings, alightings, and vehicle loads proceeds from this. Future demand could also be inferred from the use of journey planners and mobile phone applications, since the usage of these services implies (with some uncertainty) that a journey is being (or will be) made. For example, if a person uses a mobile phone to check the estimated arrival time of the next vehicle at a stop, the person might be walking toward or already waiting at the stop; the uncertainty of this inference decreases if the location of the mobile device is known. Service alerts can alter typical demand, for example by encouraging passengers to find other modes, routes, or times to make their trip when disruptions lead to overcrowding and longer than usual waiting and trip times. Demand models capturing how people react to service alerts and other types of information could be useful. Future dwell times can be obtained from estimates of current load, future boardings, and future alightings, combined with a dwell time model calibrated with historical data.

Future running times can be obtained from historical AVL data. If overtaking is not possible in all or part of the route (e.g. track, busway, or road), blocking constraints can be considered. Future vehicle locations can be predicted based on the sum of running times, time lost in stopping (especially important in bus service), and dwell times, but also considering blocking constraints and control actions. To calculate the expected location, the expected values of these elements can be added deterministically. If the distribution of locations is desired, then simulation or numerical convolution of the distributions of each factor are necessary.

1.3 Model-Based Operations Control

All these types of information can be leveraged by operators to measure and predict performance and intervene in real time with supply-side and demand-side actions. Performance can be measured in terms of passenger waiting times, in-vehicle times, and crowding, and can include operator-focused measures such as driver lateness. Supply-side interventions include control actions such as holding, short-turning, deadheading, and expressing, as well as use of spare vehicles and drivers, rerouting, and introduction of temporary services during severe disruptions. Demand-side interventions include generation of service alerts and dissemination of real-time information to passengers through web-based journey planners, social media, and Internet-enabled mobile devices. These measures can curtail demand, especially when a transit system faces significant disruptions, and can help reduce the number of people affected as well as shorten recovery time. Figure 1-1 illustrates the flow of data, information, and real-time operations planning and control decisions, much of which can be automated with machine to machine communications. This research focuses on supply-side interventions.

Service quality depends, in the first place, on an appropriate allocation of resources and careful operations planning. It is difficult to provide good service if the active fleet is too small for the running time distributions, or if the duty schedule is unrealistic. Second, reliable infrastructure and rolling stock are required. Only when these, and other, factors have been considered can high quality be delivered. While these a-priori considerations make it possible to offer high-quality service, performance deteriorates in the absence of supervision and control through the bunching mechanism in which a vehicle with a long leading headway experiences longer dwell times and the trailing vehicle catches up. The results are greater mean waiting times, increased perceived crowding, and less predictable trip times, all of which degrade service quality. Real-time control is especially important in systems operating at, or near, capacity, as they can help provide the best service possible until a long-term capacity upgrade can take place.

Effective real-time control regulates headways though control actions such as holding, enforcing boarding limits, short-turning, deadheading, and expressing. Holding is the most commonly employed of the strategies, and has been found to be the single most effective strategy in terms of total passenger waiting time reductions (Eberlein et al., 2001). It consists of extending the normal dwell time of a vehicle at a stop (or station) when its leading headway is shorter than it should be. This increases the number of passengers boarding the vehicle downstream of the holding location by waiting for more passengers to arrive at downstream stops or stations, thus preventing the vehicle's dwell times from being shorter than usual, which eventually leads to bunching. The goal is usually to hold vehicles such that they are evenly spaced (in time) across stops. Depending on the system, this may be done exclusively at terminals (dispatch headway regulation), at terminals and key stops or stations, or at any stop. Holding (empty) vehicles at terminals is not as onerous for passengers, and is critical for obtaining high service quality (Eberlein et al., 2001).

Boarding limits have the opposite effect of holding. By limiting the number of passengers that would board in an uncontrolled scenario, the dwell time of a

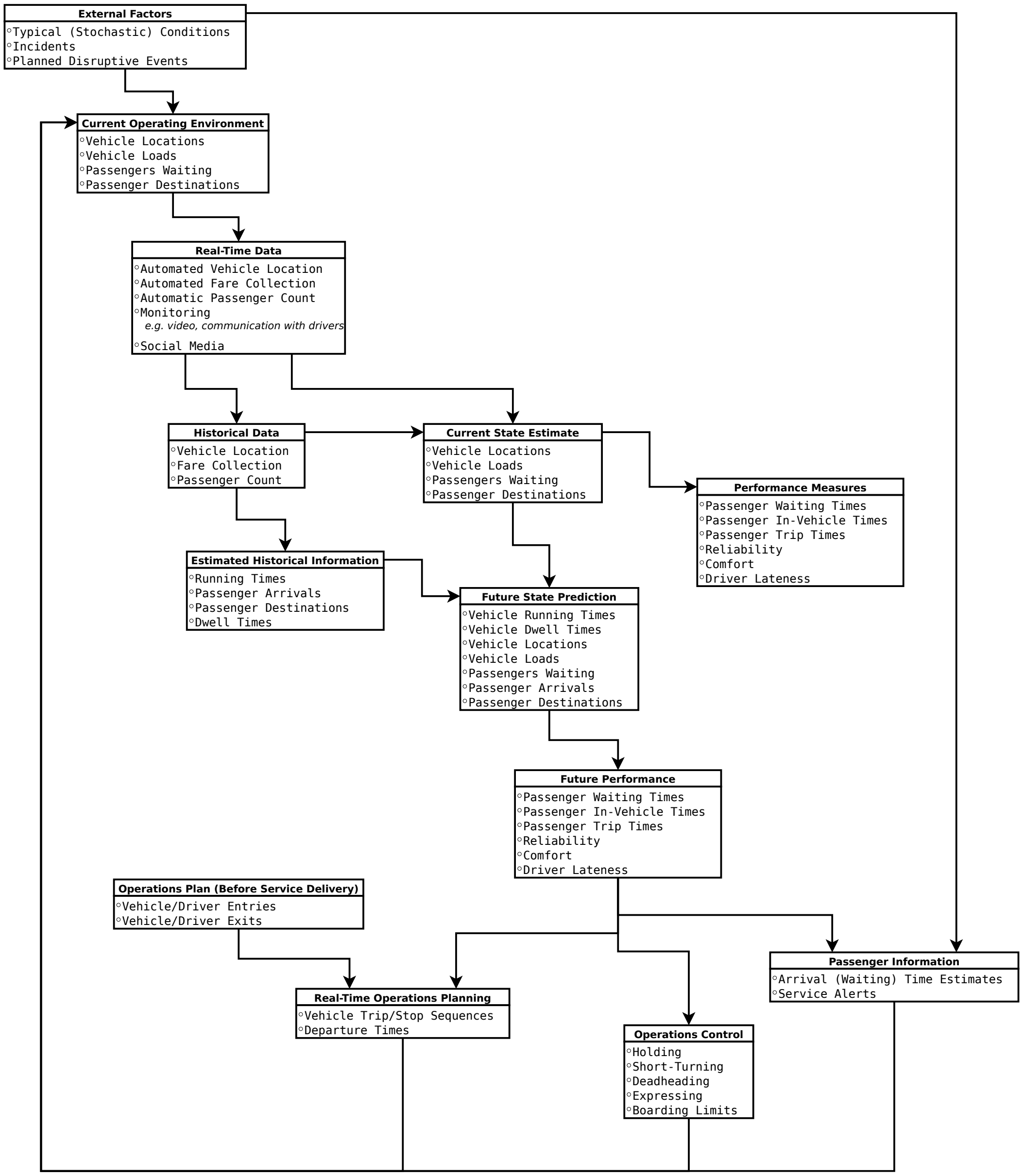


Figure 1-1: Data, Information, and Real-Time Operations Optimization

vehicle at a station or stop is shortened, and in-vehicle crowding is controlled. This allows the vehicle to move faster when it has fallen behind. (Delgado et al., 2012) Boarding limits are more common in train service than bus service, as train doors are perceived as automatic (in some cases they are), and no explanation is sought by passengers who cannot board.

Short-turning, deadheading, and expressing are more drastic strategies that change the sequence of vehicles through stop-skipping. These strategies can greatly affect service quality, but they require coordination and awareness of not just the leading headway, but the overall distribution of vehicles on the route and the number of passengers on buses and stops or stations. Hence, these strategies are best evaluated from a control center with a complete view of the system. Short-turning consists of ending a trip before reaching the downstream terminal, forcing any passengers in the vehicle to alight, and turning the vehicle to begin service in the opposite direction. It can be effective when there is a bunch in one direction and a large gap in the other, especially if one of the bunched vehicles is not highly loaded and the passengers in it are allowed to transfer to the other vehicle in the bunch (so that they do not have to wait at the stop). Short-turning can be more benign when it is decided (and announced) before beginning a trip, with no passengers in the vehicle. Deadheading consists of moving an empty vehicle not in revenue service from one location to another. For example, a bus at a terminal can be deadheaded to the middle of the route to begin revenue service there. In rail services, deadheading options are more limited, but sometimes it is possible to skip a few stations after departing the terminal, and then begin revenue service. Expressing consists of skipping stops or stations while in revenue service, forcing passengers whose destinations will be skipped to alight and continue their journey on the next vehicle (or on foot).

At least six factors influence the degree to which information improves performance: the availability and quality of data, the accuracy of information, the ability of operations planning and control models to utilize information, the sensitivity of optimal policies to information, the types of interventions available, and the operating environment.

Data The effectiveness of control decisions depends, in the first place, on the availability and quality of data. Historical and real-time data on vehicle locations and demand enable the use of models and control strategies designed for data-rich operations. The frequency and delay with which real-time data is received from connected sensors, and measurement errors at the source play a role. Sophisticated models may not provide value if vehicle locations are not received in a reliable and timely manner.

Information Modeling Information models process and combine data to obtain useful information. A stream of vehicle positions may be processed to remove outliers and determine running times. Automated passenger counting data may be processed to obtain vehicle loads. The effectiveness of control decisions depends, in part, on the amount and types of information used to model system states and optimize operations. Characterizing a transit service with more types of information can potentially improve the effectiveness of control actions by painting a more complete picture and being able to measure service quality more directly in the control objective. For example, moving from a characterization based only on vehicle locations to one combining vehicle locations and loads might enable a control framework to avoid holding full vehicles. The accuracy of the models used to extract information from data play a role. Modeling errors can lead to information errors, and possibly suboptimal or even counterproductive control actions.

Operations Planning and Control Models Model capabilities determine whether different types of information can be used and how much they can contribute to improving operations. For example, a simple even headway policy is insensitive to loads, so it would hold a full vehicle even if there is load information showing that the vehicle is full. In research, model capabilities limit the ability to quantify the value of certain types of information. For instance, the optimization models developed in this research all model running times and demand dynamically, allowing dynamics to affect service. Without the ability to model running times and demand dynamically, transit services could not benefit in real-time from information about dynamics, and it would not be possible to measure, through experiments, its potential value.

Sensitivity of Optimal Policies to Information The effectiveness of control decisions also depends on the extent to which better information leads to different (better) decisions, and thereby improved service quality. For example, capturing stochasticity of demand might not lead to a change in control policies if total arrival rates are not highly variable, or if variability in demand does not affect which control decisions are optimal, although in these cases better information could lead to a higher certainty that good decisions are being made.

Available Interventions The extent to which information improves performance depends on a transit service's operational strategy, including dispatch strategy (e.g. schedule-based vs. headway-based) and available control strategies (e.g. holding alone, short-turning alone, holding and short-turning combined). For example, some types of information might be significantly more useful to make good short-turning decisions, but not generally influence holding decisions.

Operating Environment The extent to which information changes the decisions being made might depend on system characteristics such as running times, demand patterns, and allocated resources (which govern effective cycle time reliability). For example, some kinds of information, richer models, and intervention strategies might be especially useful in near-capacity operations, but less so when vehicles do not fill up.

1.4 Research Objectives

This research aims to advance, through new frameworks and methodologies, the utilization of historical and real-time information to improve high-frequency public transit operations. It takes advantage of the trend of increasing connectivity of sensors and devices, and the ability it creates to observe the state of a transit line in real-time, to address the problem of disruptions in high-frequency public transportation services. This research focuses on supply-side interventions, and specifically in the following objectives.

1. Develop the framework and models required to capture information about dynamics of running times and demand when optimizing operations control decisions, focusing on holding control and dynamics generated by typical exogenous changes in a transit line's operating environment throughout the day. Determine to what extent, in what manner, and under which circumstances information about dynamics can improve the effectiveness of control. This objective is addressed in Chapter 2.

2. Develop the framework for controlling a transit service with information about events, building upon the model developed in Chapter 2 to optimize holding control reflecting anticipated event-driven dynamics. Investigate both foreseen and unforeseen events. Determine if controlling with information about events can improve the performance of a transit line, and whether any benefit is robust to errors of information. This objective is addressed in Chapter 3.

3. Develop the concept of, and framework for, schedule-free real-time operations planning of high-frequency transit, which would allow operations plans to benefit from real-time information. Discuss potential methodological approaches, and experimentally determine if such an approach is feasible and potentially beneficial. This objective is addressed in Chapter 4.

1.5 Dissertation Structure

The following three chapters of this dissertation present separate but related research. Chapter 2 investigates how information about dynamics in running times and demand can improve the effectiveness of holding control in high-frequency transit. Previous holding control strategies either do not forecast system states to optimize holding times or they rely on forecasts based on running times and demand assumed to be static. The chapter presents a holding optimization model that reflects dynamics, thus producing a holding policy that accounts not only for the current state of the system, but also for expected changes in running times and demand. Its effectiveness is evaluated within a simulation environment. The results show that control based on dynamic inputs outperforms its static equivalent in high demand cases where passengers may be left behind at stops and, to a lesser extent, in low to moderate demand cases with time-varying running times.

Chapter 3 explores how information about potentially disruptive events can improve the effectiveness of holding control in high-frequency transit. Some events, such as signal failures in rail transit and traffic accidents along a bus route operating in mixed traffic, cause disruptions that are unpredictable. Others, such as concerts or sport contests, cause disruptions that can be foreseen. In some cases, information can be used to predict disruptions, which can then be modeled using dynamic functions of running times and demand, using the model developed in Chapter 2. Past research has explored the use of operations control to respond to disruptions once service deteriorates. By dynamically modeling expected changes in running times and demand during events, the framework presented in Chapter 3 enables anticipatory control strategies. A holding optimization model capturing event-driven dynamics is applied to a simulated transit service experiencing disruptions induced by an unforeseen and a foreseen event. Controlling operations with awareness of events improves performance in both cases. However, erroneous estimates of the time an event will occur can lead to counterproductive control policies.

Chapter 4 introduces a schedule-free paradigm for high-frequency transit operations, in which operations are not only controlled but also planned in real-time, taking advantage of real-time information for optimizing vehicle trip sequences in a way that maximizes service quality while satisfying operator resource constraints. Previous research in operations control, including the control strategies presented in chapters 2 and 3, neglects planned driver duty end times and changes in number of active vehicles. (Strategies are usually evaluated through simulations with a constant number of active vehicles.) When these control strategies are applied to a real transit service, they optimize control actions counting on a vehicle that will be taken out of service to continue serving trips, or not reflecting the imminent entry of vehicles. They may also delay a vehicle that is late with respect to the schedule, putting desirable control outcomes at odds with schedule constraints. By reconciling the service quality improvement and operator constraint satisfaction goals, the schedule-free paradigm increases the utility of information for high-frequency transit operations. Transit services running under the schedule-free paradigm adapt to current and expected future conditions. After introducing the new paradigm conceptually, Chapter 4 develops a framework and methodology for schedule-free operations and discusses some implications. An example application demonstrates the feasibility and potential of the new paradigm.

Chapter 5 summarizes the research methodology, findings, and contributions, and discusses potential future research.

The optimization problems in this research are complex, and their objective functions are, generally, not globally convex, so there may be multiple local optima. Throughout the dissertation, the term *optimal* refers to the lowest cost solution found by the optimization algorithm.

Chapter 2

Holding Control with Dynamics

Operations control is an important means of improving service quality in high-frequency public transport systems. It is based on continuous monitoring of the system and supply-side interventions with the aim of providing the best possible service to passengers with the available resources. Holding, the most commonly employed intervention, consists of intentionally delaying a vehicle, possibly at the expense of extending trip times for passengers on board, in order to reduce the waiting time of passengers who will board downstream. Previous research has established that holding is generally the single most effective type of intervention, and a number of methods to determine holding policies have been proposed and evaluated, ranging from simple heuristics to sophisticated model-based optimization. (Eberlein et al., 2001)

Past research has focused on evaluating the effectiveness of different control strategies, largely based on static assumptions with respect to both running time and demand. Within the family of real-time control strategies based on future system state prediction models, none to date have modeled running times and demand dynamically. Holding strategies based on static running times and demand respond to disruptions in the initial system state. For example, if a vehicle has been delayed, holding can help prevent bunching from occurring. However, these strategies cannot anticipate systematic changes in the operating environment and preemptively consider those changes in the production of a control plan. A strategy that models running times and demand dynamically can do this. For example, such a strategy may be able to anticipate how additional demand leading into the rush hour may cause disruptions, and therefore produce a control plan that recognizes that such disruptions may develop.

This has significant implications in the realm of real-time information and operations control. By being able to react not only to what is already known in the initial system state but also to what can be anticipated based on historical experience, the model proposed in this research could potentially improve performance beyond that achieved with previously developed control strategies. Future running times and demand could be predicted based on real-time information of current conditions in addition to historical or typical conditions.

For instance, an unusually high demand observed in some part of the network might be used to infer additional demand in another part of the network due to transfers. Or an accident might be reported at some intersection, allowing an operator to anticipate traffic delays along a corridor before the congestion develops. The aim of this research is to develop and test a model that can be used in these dynamic contexts, focusing on cases where future running times and demand are known in advance.

The remainder of this chapter is organized as follows: Section 2.1 reviews the literature, Section 2.2 presents the framework and mathematical model, Section 2.3 discusses the implementation of this model within a simulation model for evaluation purposes, Section 2.4 presents and discusses the results from the application of the control model, and Section 2.5 draws conclusions.

2.1 Literature Review

Early research on the holding problem did not consider the availability of real-time information (see Osuna and Newell, 1972, Barnett, 1974, Newell, 1974, Turnquist and Blume, 1980, and Abkowitz and Lepofsky, 1990). Since then researchers have proposed a number of control strategies with control actions based on both typical system characteristics (e.g. running times and demand) and real-time information (e.g. vehicle locations). Holding strategies differ in objectives, underlying models and solution methods, and information utilized. Common objectives include schedule adherence (Adamski and Turnau, 1998), headway adherence (Rossetti and Turitto, 1998), headway regularity (Daganzo, 2009 and Bartholdi and Eisenstein, 2012), and cost minimization (Eberlein et al., 2001, Delgado et al., 2009, Delgado et al., 2012, and Sáez et al., 2012). Schedule adherence is a suitable objective for long headway service, while headway regularity is a suitable objective for short headway service. Strategies based on cost minimization employ mathematical programming or other optimization methods and a variety of information including vehicle locations, loads, and passenger arrival rates, sometimes in a rolling horizon formulation. In some cases holding is combined with other control strategies such as short-turning (Shen and Wilson, 2001), boarding limits (Delgado et al., 2009 and Delgado et al., 2012), and signal priority (Chandrasekar et al., 2002).

Control strategies are typically evaluated through simulation. Performance is measured in terms of passenger waiting times and trip times (expected values and variability), often complemented with measures of headway regularity, loads, etc. Researchers have been effective at using this evaluation framework to demonstrate how the performance of transit services can improve with enhanced availability and utilization of information, as well as innovative prediction and control models.

Previous research has not addressed the value of modeling system characteristics dynamically. While many of the recently proposed strategies utilize real-time information, very few are able to consider predictions of future system states involving dynamic running times and demand. The few with that (limited) ability have been tested in simulation environments having time-independent running times and demand. For example, see Dessouky et al. (2003), Daganzo (2009), and Sáez et al. (2012). In other cases, the simulation environment features time-dependent running times and demand, but the strategies used cannot model the dynamics and instead assume typical period-level constants.

Adamski and Turnau (1998) develop a set of holding control strategies based on control theory. The aim is primarily punctuality, but the authors suggest a variation that adapts the punctuality control procedure to achieve headway regularity.

Rossetti and Turitto (1998) develop a holding strategy based on a dynamic threshold. The threshold used to determine the desired preceding headway, and therefore holding time, of a vehicle is chosen from a range of thresholds including the scheduled headway, with the aim of reducing holding time, and hence in-vehicle time added by holding. The term *dynamic* refers to the threshold and not to the running time and demand information.

Eberlein et al. (2001) formulate the holding problem as a deterministic quadratic program in a rolling horizon scheme, allowing real-time information to be taken into account. Their model includes the effect of dwell time on vehicle delay and headways, and the optimization objective is to minimize total passenger waiting times. Running times between stations and passenger arrival rates are assumed constant over the rolling horizon. The formulation includes a constraint that prevents late vehicles from being held. The resulting program is non-convex quadratic. The researchers find that holding policies are mainly sensitive to vehicle headway patterns and much less sensitive to passenger demand patterns, and that the impact of holding a vehicle on the trajectories of vehicles upstream diminishes quickly.

Shen and Wilson (2001) formulate a mixed integer program for holding, short-turning, and expressing trains on an urban rail line in the case of minor disruptions. Passenger demand and running times between stations are treated as constants.

Chandrasekar et al. (2002) test the strategy of regulating bus spacing by simultaneously providing signal priority to the leading vehicle and holding the trailing vehicle when the space between the vehicles shortens. Since the control strategy is reacting to deviations from the target vehicle spacing, it is not able to produce control policies based on dynamic running times and demand.

Zhao et al. (2003) present a distributed control approach in which vehicles and stops act as agents that communicate in real-time to coordinate departure times of vehicles from stops. They analyze the performance of their strategy, and other strategies, using simulation under a variety of conditions (including long headway scheduled service), including random bursts of passenger arrivals. Although they assess the performance of the model under dynamic conditions, their strategy does not consider expected dynamics.

Dessouky et al. (2003) compare control strategies that depend on communication, tracking, and passenger counting technologies and those using only local information. The application of interest is schedule coordination at a terminal. Among the strategies considered, some make use of predicted arrival times at the terminal and loads, which makes it possible to account for dynamic running times and demand. However, the researchers do not focus on the value of capturing predicted dynamics from the operations control perspective. Their simulation model uses time-independent running time distributions and mean passenger arrival rates.

Sun and Hickman (2008) formulate a convex quadratic program with linear constraints for optimizing holding times of vehicles at multiple stations. They assume a constant passenger arrival rate and vehicle travel time between adjacent stops.

Puong and Wilson (2008) develop a real-time disruption response model for rail transit focusing on holding. The model is a non-linear mixed integer program and captures passengers left behind at stations. The researchers assume constant passenger arrival rates and dwell times.

Delgado et al. (2009) formulate a non-convex quadratic program with linear constraints for holding and boarding limits. The model captures vehicle capacity constraints, and can be used in a rolling horizon optimization application. Passenger arrival rates per stop and travel times between stops are assumed deterministic, known, and constant over the optimization horizon.

Daganzo (2009) develops a holding strategy in which vehicles are analyzed in pairs. The trailing vehicle of a vehicle pair is delayed by holding when the headway shortens, and instructed to speed up (perhaps through denied boardings) when the headway lengthens. The authors state that the model can be extended to time-dependent demand and running times by using run-specific parameters and making adjustments to average inter-stop running times. However, they do not test this, and the method, which looks only at vehicle pairs, is not able to capture dynamics of demand or running times happening farther in the future.

Yu and Yang (2009) propose a two-step holding strategy. In the first step, a support vector machine is used to predict whether or not a vehicle will depart early from the next stop if it is held at the current stop, possibly considering running time dynamics. Holding is only considered if a vehicle is early now and is also predicted to be early at its next stop. In the second step, a genetic algorithm is used to minimize a combination of waiting cost and in-vehicle cost. Passenger arrival rates are modeled as time-independent, and the dynamics of running time predictions are limited to the next stop. The simulation experiment shows that running times predicted by the support vector machine are more accurate than mean running times, and that their optimization-based strategy reduces passenger cost more than a schedule-based strategy.

Xuan et al. (2011) develop a holding control strategy based on a virtual schedule. A one parameter version of the method can be optimized in closed form and is shown to be near-optimal and to outperform other holding strategies. Mean running times between stations are modeled as time-independent constants and the passenger arrival process is assumed stationary.

Daganzo and Pilachowski (2011) develop a holding control model based on control theory. Headways are adjusted through holding considering the preceding and following headways. In their continuum idealization, the passenger arrival process is spatially homogeneous and time-independent. Vehicles are assumed to run at a fixed average speed between stops.

Cats et al. (2011) compare the effectiveness of holding based on schedule adherence, target headway, and even headway strategies. They find that the even headway strategy leads to the best performance in terms of headway regularity, trip time savings, and schedule adherence at a relief point. The strategies they consider are local and myopic (i.e. they do not involve prediction or optimization). Hence, they do not yield policies sensitive to time-dependent running times or demand.

Bartholdi and Eisenstein (2012) develop a strategy that holds vehicles based on their following headway, and show that it leads to even headways. The strategy can react to perturbations but does not take into account predicted future changes in running times or demand.

Delgado et al. (2012) formulate a non-linear model for optimization of holding times and metering of boardings of all vehicles of a transit line at all stops, taking into account vehicle capacity constraints. The model, which this research builds upon, can be used in a rolling horizon framework. The authors show its effectiveness in simulation experiments. The model assumes time-independent passenger arrival rates and running times between stops.

Sáez et al. (2012) propose a control strategy that models demand stochastically and involves a discrete-time event-based predictive model. The strategy is applied in a rolling horizon framework and can suggest holding and expressing control actions. A genetic algorithm is used to identify control actions in reasonable computation times. The authors test the strategy in simulations with constant vehicle speed and Poisson demand having constant arrival rates by origin-destination pair. Muñoz et al. (2013) compare the holding strategy of this paper to that of Delgado et al. (2012).

Chen et al. (2013) investigate the strategy of holding a group of buses at one or more control points, considering boardings while holding. Their formulation uses constant deterministic running times between stops and passenger arrival rates.

This research extends the literature of real-time control for high-frequency transit services by presenting a model that explicitly incorporates the dynamic nature of running times and demand. The following sections present the proposed model framework and evaluate its performance.

2.2 Framework and Formulation

In this section we present the framework and formulation of a deterministic rolling-horizon performance model of a high-frequency transit service. This model extends the work of Delgado et al. (2012). It takes as inputs dynamic running time and demand functions, current system state (e.g. vehicle positions, load estimates, and estimated number of passengers currently waiting at stops), and a set of planned holding times for each vehicle at each stop. It outputs predicted future states (e.g. departure times and loads) through the next cycle. Demand is modeled with time-varying arrival rates at the origin-destination pair level. Vehicle movement is modeled with time-varying running times.

The model is used to optimize holding times at any stop where control can be applied. Figure 2-1 illustrates how the different model components interact. The framework consists of an optimization model with two main components: a performance model and a cost model. The optimization model feeds the performance model with dynamic running time and demand functions, the current system state, and candidate holding times. The performance model uses these inputs to predict how the system will evolve, including vehicle arrival and departure times, boardings and alightings, passengers left behind, and loads for all vehicles at all stops. The prediction is passed to the cost model, which gives a scalar mean cost per passenger reflecting waiting times and in-vehicle delay due to holding. The optimization model considers the costs of previously evaluated candidate solutions to select new candidate holding times, until a (local) minimum cost is found. The optimization, performance, and cost models are described in Sections 2.2.2, 2.2.3, and 2.2.4, respectively.

All of this takes place in the context of a rolling prediction horizon, which is defined (as in Delgado et al., 2012) to cover the departure of each vehicle from every stop once, starting with the next stop to be visited and finishing with the previously visited stop, to complete a cycle. This is illustrated in Figure 2-2 for two vehicles. Since the horizon boundary is defined spatially (i.e. a fixed number of stops), each vehicle's last stop visit in the horizon may happen at a different time.

The performance model can be used to make a prediction of how the system will evolve outside the optimization context. This is useful to estimate future arrival and departure times at stops considering time-dependent running times and demand.

2.2.1 Assumptions

The following assumptions are made:

1. The model is deterministic. Stochasticity, which is the very phenomenon that leads to the need for operations control, is neglected. This is done in the interest of tractability. We hypothesize that the consequences

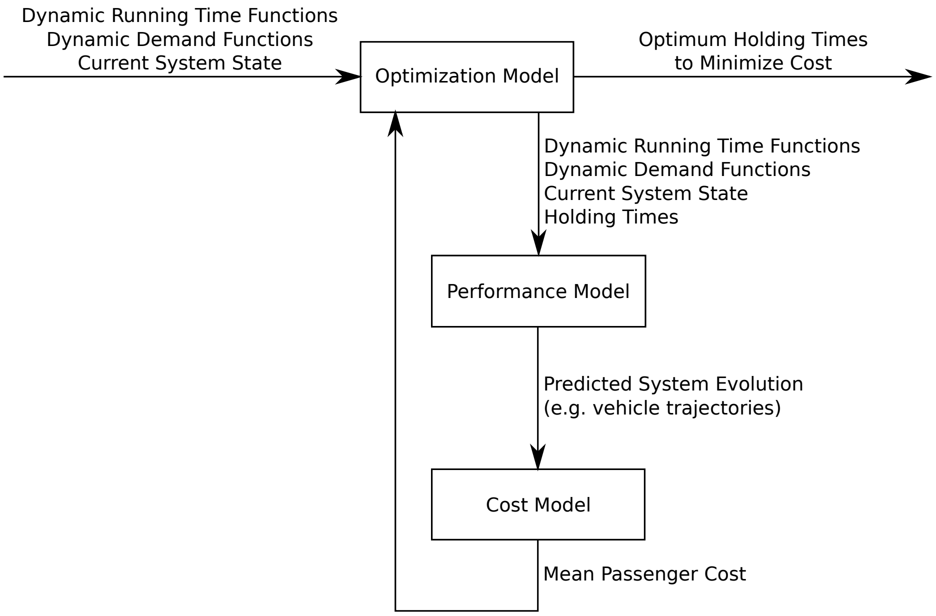


Figure 2-1: Optimization Framework

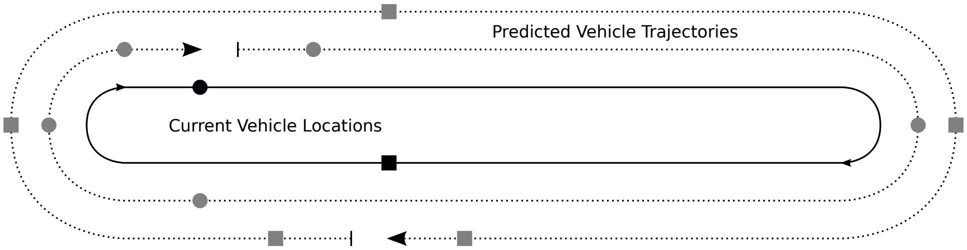


Figure 2-2: Rolling Prediction Horizon

of ignoring stochasticity should not greatly affect the early parts of the model horizon, which are the most relevant for control decisions.

2. Passenger demand is modeled with continuous variables. This allows a simplified model of dwell times and avoids modeling discrete events such as the boarding of an individual passenger.

3. Vehicles stop at every stop. This assumption avoids decision variables concerning the conditions under which a vehicle would skip a stop. This assumption is accurate for rail services, but may have an effect in bus services if many stops are skipped and a significant portion of running times is deceleration and acceleration times at stops.

4. Vehicle order is preserved over the prediction horizon. This assumption can be relaxed with additional bookkeeping not presented in this chapter in the interest of clarity. A relaxed version of the formulation could be used for bus services in which overtaking happens frequently or when the insertion and removal of vehicles throughout the day is to be modeled. (The event-based model presented in Chapter 4 captures vehicle entries, exits, and reordering.)

These assumptions concern the model, and do not necessarily apply to the simulation used to evaluate the effectiveness of the proposed control strategy.

2.2.2 Holding Time Optimization

The performance and cost models can be used to predict vehicle trajectories and passenger costs given a particular set of holding times for each vehicle and stop in a service with n_V vehicles and n_S stops. Treating them as a mathematical function $f : \mathbb{R}^{n_V n_S} \mapsto \mathbb{R}$ mapping a set of holding times $h_{v,s}$ to a scalar cost, they can also be used to obtain a set of optimal holding times in an optimization context. The objective is to minimize mean cost per passenger over the prediction horizon, which combines waiting time and in-vehicle delay, subject to constraints on holding times at stops:

$$\begin{aligned}
\underset{h_{v,s}\ \forall v \in V\ \forall s \in S}{\text{minimize}} \quad & \frac{W_V + \theta_S W_S}{P} && (2.1)\\
\text{subject to} \quad & \text{vehicle movement constraints} && (2.2)\\
& \text{passenger activity constraints} && (2.3)\\
& 0 \leq h_{v,s} \leq h_s^{\max} \quad \forall v \in V \quad \forall s \in S && (2.4)
\end{aligned}$$

where W_V and W_S are the total in-vehicle delay and waiting time for all passengers in the prediction horizon, P is the total number of passengers, and θ_S determines the relative disutility of waiting at a bus stop with respect to delay inside a vehicle.

Constraints (2.2) and (2.3) are handled by the performance model presented in Section 2.2.3. Constraint (2.4) states that holding times must be non-negative and may not exceed the maximum allowed by policy. For example, a maximum holding time of two minutes might be set at stops where holding is allowed, while longer maximums can be set at terminals having designated space for vehicles to hold without interfering with other vehicles. The upper bound can be made very large, effectively removing this constraint, and it can also be set to zero, effectively preventing holding at a particular location.

Section 2.2.4 presents the mathematical formulation of the cost model. The objective function is not globally convex, so there may be multiple local optima. In this research, the term *optimal* refers to the lowest cost solution found by the optimization algorithm.

2.2.3 Performance Model

The performance model predicts how the system will evolve over the rolling prediction horizon. The prediction includes vehicle arrival and departure times, running times between stops, boardings and alightings, dwell times, and number of passengers left behind due to capacity constraints. The prediction of the future states of the system is used by the cost model to calculate the cost associated with a candidate set of holding times.

Stops are labeled in order from 1 to n_S and vehicles from 1 to n_V, with stop n_S and vehicle 1 being the furthest downstream stop and vehicle, respectively. The formulation uses rolling addition and subtraction for vehicle and stop indexes. For example, if $v = 1$, $v - 1 = n_V$, and if $s = n_S$, $s + 1 = 1$. Greek letters are used to represent passenger activity. Primed variables (e.g. d') represent known quantities observable before the current time t_0. Barred variables (e.g. $\bar{d}$) represent constrained quantities.

The arrival time $a_{v,s}$ of a vehicle v at its first stop in the rolling horizon (stop $s = e_v + 1$) is obtained by adding its running time to the departure time from the previously visited stop. This departure time d'_{v,e_v} is a constant that was observed before t_0. The unconstrained running time is determined by evaluating the dynamic running time function $r_{e_v}(t)$ of the vehicle's initial stop (stop e_v) at the vehicle's departure time.

$$a_{v,s} = d'_{v,e_v} + r_{e_v}\left(d'_{v,e_v}\right) \quad \forall v \in V \quad s = e_v + 1 \tag{2.5}$$

The model requires preservation of vehicle order throughout the optimization horizon. Extra running time $\bar{r}'_v$ can be added to prevent a vehicle from arriving at a stop before the preceding vehicle. The preceding vehicle's arrival time is known because it occurs before t_0.

$$\bar{r}'_v = \max\left(0, a'_{v-1,s} - a_{v,s}\right) \quad \forall v \in V \quad s = e_v + 1 \tag{2.6}$$

The (possibly) constrained arrival time $\bar{a}_{v,s}$ at a vehicle's first stop is the unconstrained arrival time plus any extra running time required to prevent overtaking.

$$\bar{a}_{v,s} = a_{v,s} + \bar{r}'_v \quad \forall v \in V \quad s = e_v + 1 \tag{2.7}$$

These terms collectively represent an estimate of a vehicle's arrival time at its first stop in the optimization horizon. Unconstrained arrival times at subsequent stops are obtained from the addition of running time to the departure time from the previous stop, $d_{v,s-1}$. Running times $r_s\left(t\right)$ are dynamic functions.

$$a_{v,s} = d_{v,s-1} + r_{s-1}\left(d_{v,s-1}\right) \quad \forall v \in V \quad \forall s \in \left[e_v + 2, e_v\right] \tag{2.8}$$

Extra running time may have to be added to prevent overtaking. This applies starting at the first stop that a vehicle visits following a visit by the preceding vehicle at the same stop in the optimization horizon. An exception is made for the first vehicle (i.e. the one furthest downstream) in the special case that all vehicles are between the same two stops at the beginning of the optimization horizon; otherwise the model could be infeasible.

$$\bar{r}_{v,s-1} = \begin{cases} \max\left(0, \bar{a}_{v-1,s} - a_{v,s}\right) & \forall v \in V \quad \forall s \in \left[e_{v-1} + 1, e_v\right] \\ & \text{unless } v = 1 \text{ and } e_1 = e_{n_V} \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

The (possibly) constrained arrival time is the unconstrained arrival time plus any extra running time required to prevent overtaking.

$$\bar{a}_{v,s} = a_{v,s} + \bar{r}_{v,s-1} \quad \forall v \in V \quad \forall s \in \left[e_v + 2, e_v\right] \tag{2.10}$$

Unconstrained departure times $d_{v,s}$ are obtained by adding dwell time $\delta_{v,s}$ and holding time $h_{v,s}$ to arrival times.

$$d_{v,s} = \bar{a}_{v,s} + \delta_{v,s} + h_{v,s} \quad \forall v \in V \quad \forall s \in S \tag{2.11}$$

In order to prevent overtaking, it is necessary to delay a vehicle's departure by $\bar{h}_{v,s}$ when the preceding vehicle has not yet departed the same stop. As with arrival times, the overtaking constraint starts to apply at the first stop that a vehicle visits following a previous visit in the optimization horizon, and an exception is made for the first vehicle in the special case that all vehicles are between the same two stops at the beginning of the optimization horizon.

$$\bar{h}_{v,s} = \begin{cases} \max\left(0, \bar{d}_{v-1,s} - d_{v,s}\right) & \forall v \in V \quad \forall s \in \left[e_{v-1} + 1, e_v\right] \\ & \text{unless } v = 1 \text{ and } e_1 = e_{n_V} \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

The (possibly) constrained departure time is the unconstrained departure time plus any extra non-control holding time (i.e. blocked time) required to prevent overtaking.

$$\bar{d}_{v,s} = d_{v,s} + \bar{h}_{v,s} \quad \forall v \in V \quad \forall s \in S \tag{2.13}$$

A vehicle's dwell time is the greater of the boarding and alighting times. Boarding and alighting times are modeled as linear functions of the number of boarding $\bar{\beta}_{v,s}$ and alighting $\alpha_{v,s}$ passengers, respectively. Constant boarding and alighting times per passenger, τ_b and τ_a, are assumed.

$$\delta_{v,s} = \max\left(\tau_b \bar{\beta}_{v,s}, \tau_a \alpha_{v,s}\right) \quad \forall v \in V \quad \forall s \in S \tag{2.14}$$

Demand is modeled at the origin-destination pair level (s_b to s_a). For the first visit to each stop, it includes passengers who are already at the stop at the beginning of the optimization horizon, ι'_{s_b,s_a}. For subsequent visits, it includes passengers who were unable to board the preceding vehicle because it was full, π_{v-1,s_b,s_a}. The number of arriving passengers is determined by integrating time-dependent passenger arrival rate functions $\lambda_{s_b,s_a}(t)$. The number of passengers waiting to board per origin-destination pair is given by

$$\beta_{v,s_b,s_a} = \begin{cases} \iota'_{s_b,s_a} + \displaystyle\int_{t_0}^{d_{v,s_b}} \lambda_{s_b,s_a}(t)\,\mathrm{d}t & v = g_{s_b} \quad \forall s_b \in S \quad \forall s_a \in S \\ \pi_{v-1,s_b,s_a} + \displaystyle\int_{d_{v-1,s_b}}^{d_{v,s_b}} \lambda_{s_b,s_a}(t)\,\mathrm{d}t & \forall v \in V \quad \forall s_b \in [e_{v-1}+1, e_v] \\ & \forall s_a \in S \end{cases} \tag{2.15}$$

where vehicle g_{s_b} is the first to visit stop s_b in the horizon. The total number of waiting passengers is

$$\beta_{v,s} = \sum_{s_a \in S} \beta_{v,s,s_a} \quad \forall v \in V \quad \forall s \in S \tag{2.16}$$

The actual number of boardings cannot exceed the remaining capacity, which is determined by subtracting load $l_{v,s}$ and adding alightings to capacity k.

$$\bar{\beta}_{v,s} = \min\left(\beta_{v,s}, k - l_{v,s} + \alpha_{v,s}\right) \quad \forall v \in V \quad \forall s \in S \tag{2.17}$$

When the number of passengers who want to board exceeds the remaining capacity, it is assumed that all passengers, regardless of destination, are equally likely to board.

$$\bar{\beta}_{v,s_b,s_a} = \frac{\bar{\beta}_{v,s_b}}{\beta_{v,s_b}} \beta_{v,s_b,s_a} \quad \forall v \in V \quad \forall s_b \in S \quad \forall s_a \in S \tag{2.18}$$

The number of passengers π_{v,s_b,s_a} prevented from boarding (by origin-destination pair) is the difference between passengers waiting and those who board.

$$\pi_{v,s_b,s_a} = \beta_{v,s_b,s_a} - \bar{\beta}_{v,s_b,s_a} \quad \forall v \in V \quad \forall s_b \in S \quad \forall s_a \in S \tag{2.19}$$

The number of passengers prevented from boarding vehicle v at stop s is

$$\pi_{v,s} = \beta_{v,s} - \bar{\beta}_{v,s} = \sum_{s_a \in S} \pi_{v,s,s_a} \quad \forall v \in V \quad \forall s \in S \tag{2.20}$$

Alightings come not only from passengers who board during the optimization horizon but also from passengers who are already in the vehicle at the beginning of the optimization horizon. We assume that all passengers in a vehicle at the last stop alight. Thus, the number of alighting passengers includes passengers initially in the vehicle starting at the first stop the vehicle visits and up to the last stop. We assume that there is no demand for travel from a stop to the same stop, i.e. $\beta_{v,s,s} = 0$ for every vehicle and stop.

$$\alpha_{v,s_a} = \begin{cases} l'_{v,s_a} + \displaystyle\sum_{s_b=e_v+1}^{s_a} \bar{\beta}_{v,s_b,s_a} & \forall v \in V \quad \forall s_a \in [e_v+1, n_S] \\ \displaystyle\sum_{s_b=e_v+1}^{s_a} \bar{\beta}_{v,s_b,s_a} & \forall v \in V \quad \forall s_a \in [1, e_v] \end{cases} \tag{2.21}$$

Vehicle loads $l_{v,s}$ upon arrival at the first stop are known as part of the initial state. Vehicle loads upon arrival at subsequent stops are obtained recursively.

$$l_{v,s} = \begin{cases} \displaystyle\sum_{s_a \in S} l'_{v,s_a} & \forall v \in V \quad s = e_v + 1 \\ l_{v,s-1} + \bar{\beta}_{v,s-1} - \alpha_{v,s-1} & \forall v \in V \quad \forall s \in [e_v+2, e_v] \end{cases} \tag{2.22}$$

The set of equations in this section can be evaluated recursively to obtain a deterministic forecast of how the system will evolve over the next cycle. Specifically, the arrival time, arrival load, and departure time of each vehicle are estimated for each vehicle-stop combination.

The number of passengers inside vehicles and at stops (by origin-destination pair) at the beginning of the prediction horizon (denoted by l'_{v,s_a} and ι'_{s_b,s_a}, respectively) are estimated from vehicle passage times before t_0 and dynamic demand functions. Whenever a vehicle visits a stop, the estimated number of passengers originating at the current stop and destined for downstream stops is determined using the dynamic passenger arrival rate and the time since the previous vehicle departure, while the estimated number of passengers destined for the current stop is set to zero (because they have alighted). The initial number of passengers at stops includes the estimated number of passengers previously left behind by full vehicles.

2.2.4 Cost Model

The cost model determines mean cost per passenger using the prediction information from the performance model. Cost is based on waiting time and in-vehicle delay due to holding.

In-vehicle delay due to holding is the product of the number of through-passengers in the vehicle and the holding time. The total in-vehicle delay over the horizon is given by

$$W_V = \sum_{v\in V}\sum_{s\in S}\left(l_{v,s} - \alpha_{v,s}\right)h_{v,s} \tag{2.23}$$

Waiting time has four components corresponding to whom is affected: passengers who arrive before the start of the prediction horizon, passengers who arrive after the start of the prediction horizon and are served by the first vehicle to arrive at the stop in the prediction horizon, passengers who arrive between successive vehicle visits within the prediction horizon, and passengers who cannot board the first vehicle to visit their origin stop after their arrival because it was full.

$$W_S = W_{S_0} + W_{S_1} + W_{S_2} + W_{S_3} \tag{2.24}$$

For passengers who arrived before the start of the prediction horizon, we consider only their waiting time within the prediction horizon, given by

$$W_{S_0} = \sum_{s_b\in S}\left[\left(\sum_{s_a\in S}\iota'_{s_b,s_a}\right)\left(d_{g_{s_b},s_b} - t_0\right)\right] \tag{2.25}$$

where $d_{g_{s_b},s_b}$ is the time of the first departure from stop s_b in the prediction horizon. For passengers arriving between t_0 and the first departure, waiting time depends on the arrival rate of passengers.

$$W_{S_1} = \sum_{s_b\in S}\sum_{s_a\in S}\int_{t_0}^{d_{g_{s_b},s_b}}\lambda_{s_b,s_a}(t)\left(d_{g_{s_b},s_b} - t\right)\,\mathrm{d}t \tag{2.26}$$

The waiting time of passengers who arrive in headways contained in the prediction horizon depends on both passenger arrival rates and the departure times that determine the headway.

$$W_{S_2} = \sum_{v\in V}\sum_{s_b=e_{v-1}+1}^{e_v}\sum_{s_a\in S}\int_{d_{v-1,s_b}}^{d_{v,s_b}}\lambda_{s_b,s_a}(t)\left(d_{v,s_b} - t\right)\,\mathrm{d}t \tag{2.27}$$

Some passengers find that the first vehicle to arrive at their stop is full. They are unable to board that vehicle and must wait for the next vehicle with available space. Their additional waiting time is given by

$$W_{S_3} = \sum_{v\in V}\sum_{s_b=e_{v-1}+1}^{e_v}\pi_{v-1,s_b}\left(d_{v,s_b} - d_{v-1,s_b}\right) \tag{2.28}$$

The total number of passengers who board over the entire prediction horizon is given by

$$P = \sum_{v \in V} \sum_{s \in S} \bar{\beta}_{v,s} \tag{2.29}$$

2.2.5 Control Strategy

Every time a vehicle arrives at a control point (i.e. a terminal or stop at which holding is allowed), the model presented in this section is used to optimize holding times for all vehicles at all stops over the rolling prediction horizon. (An alternative approach not used in this research is to optimize holding times periodically, e.g. every 5 minutes, and use the most recently optimized holding times for all vehicles.) A general purpose non-linear optimizer (the *Apache Commons Math* implementation of the BOBYQA algorithm by Powell, 2009) is used. The optimizer calls the performance and cost models to evaluate the objective function. The holding time suggested by the optimizer for the vehicle that triggered the modeling-optimization event at its current stop is implemented.

2.2.6 Dimensionality Reduction

The complexity of the optimization problem is a function of the product of the number of control points and the number of controlled vehicles. Since the optimization model is called frequently, it is important to control the problem size. Optimization times may become prohibitive for instances with many control points and a large fleet. We have had success implementing optimization-based control in such cases by reducing the number of stops that are control points (e.g. control points every 5 stops) and also optimizing the holding times of a control set $V_C \subset V$ of 5 vehicles, consisting of the vehicle that triggered the optimization event, the two preceding vehicles, and the two following vehicles.

When a reduced vehicle control set is used, optimization policy must be approximated for vehicles not in the control set; otherwise optimization will produce plans assuming that other vehicles are not controlled, which could be inconsistent with what would happen at a future optimization event that does hold those vehicles. Past research (e.g. see Eberlein et al., 2001), as well as our own experiments, show that optimization-based strategies usually lead to even headways. We have had good results using the even headway strategy (presented in Section 2.3) to approximate the optimization policy for vehicles not in the control set. It is also possible to use previously optimized holding times if they are available.

In the case study presented in Section 2.3, there are 8 control points and 10 vehicles, which makes 80 decision variables. This problem is realistic in size. Allowing holding at a reduced set of stops (8 out of 40) reduces the problem size sufficiently that there is no need to also reduce the set of controlled vehicles.

2.3 Application

The proposed framework is evaluated in a simulation of high-frequency bus service. This allows the comparison of different strategies in a controlled manner, where differences in performance between cases having a particular set of demand and running time dynamics are mostly due to differences in control strategies. The simulation model is event-based. Within each replication, events such as passenger and vehicle arrivals at stops and boarding and alighting activity are processed chronologically. (Sánchez-Martínez, 2012)

The simulated transit service is a simple route with 20 stops in each direction. Vehicles have capacity for 60 passengers. Boarding and alighting times per passenger are deterministic with $\tau_b = \tau_a = 2$ seconds, and the boarding and alighting processes happen in parallel. Service operates with 10 vehicles. Stochastic running times between stops are drawn from a log-normal distribution with a mean of 60 seconds and a coefficient of variation of 0.4, though the distribution is shifted by dynamics as described later.

The passenger arrival process is Poisson, with the same mean arrival rate for all origin-destination pairs in each direction. This demand specification results in (for each direction) the first stop having the highest number of passengers boarding, the last stop having the highest number of passengers alighting, and the middle stop having the highest vehicle loads. Two demand levels are considered: low crowding and high crowding. The base mean arrival rate in the first direction is set such that (if vehicles arrive every 5 minutes) peak loads reach 75% of capacity in high crowding cases and 25% of capacity in low crowding cases. The passenger arrival rates in the second direction are half those in the first direction.

The analysis period begins two hours into the simulation and lasts two hours. All passengers who arrive at their origin stop during the analysis period are included when calculating performance measures, even if their trips end after the analysis period.

We test the effectiveness of the dynamic control strategy in six different cases involving different dynamics and crowding levels, as shown in Table 2.1. There are low and high crowding variations of dynamic running times and demand, dynamic running times but static demand, and static running times but dynamic demand.

Table 2.1: Cases

Running Times	Demand	Crowding	Optimal Target Headway (min)
dynamic	dynamic	high	4.5
dynamic	dynamic	low	4.6
dynamic	static	high	4.7
dynamic	static	low	4.6
static	dynamic	high	4.5
static	dynamic	low	4.1

Dynamics in running times and demand are introduced by a dynamic factor $\phi(s,t)$ used to transform base running times and mean passenger arrival rates in the first direction. The factor is defined as follows:

$$\phi(s,t) = \begin{cases} 1.0 + 2.0\Big(t - r^*(s-1) - 2.0\Big) & \text{for } 2.0 \leq t - r^*(s-1) < 2.5 \\ 2.0 - 2.0\Big(t - r^*(s-1) - 2.5\Big) & \text{for } 2.5 \leq t - r^*(s-1) < 3.0 \\ 1.0 & \text{otherwise} \end{cases} \tag{2.30}$$

where t, time, and r^*, running time between stops, are expressed in hours. The $r^*(s-1)$ term introduces a lag for stops after the first. The dynamic transformation shifts running time distributions without affecting their variance. In cases having dynamic demand, the dynamic transformation affects the mean arrival rates $\lambda_{o,d}(t)$ governing the Poisson process. In the two cases having dynamic demand with high crowding, vehicles reach capacity and some passengers are unable to board the first vehicle.

Using this setup, we independently simulate service with the static and dynamic control using the optimization model formulated in Section 2.2. In addition to the static and dynamic optimization-based strategies, we also evaluate service with two heuristic control strategies in order to have a reference for the performance of the optimization-based strategies.

The following is a list of the strategies considered:

TH (Target Headway) Hold vehicle v at control point s to ensure preceding headways are never less than a prescribed target headway H.

$$h_{v,s} = \max\Big(0, H - \Big(d_{v,s} - d_{v-1,s}\Big)\Big) \tag{2.31}$$

where $d_{v,s}$ is the departure time of the vehicle being controlled if no holding is applied, and $d_{v-1,s}$ is the time of the previous departure from the control point. Optimal target headways for each case are found by simulating service with a range of target headways and selecting the ones yielding lowest mean passenger cost for the analysis period. Table 2.1 shows optimal target headways for each case. This control strategy gives an upper bound on the performance (lower bound on the cost) that can be attained with a constant target headway strategy, though it might be possible to improve performance further with time-varying target headways. Since the other strategies are evaluated by comparison to this strategy, the optimization of target headways reduces the reported effectiveness of the other strategies. This approach differs from the one more commonly followed in the literature (e.g. Delgado et al., 2012), in which a non-optimized schedule headway (or in some cases no control at all) is used as a base case that favors proposed control strategies.

EH (Even Headway) Hold with the aim of equalizing the preceding and following headways. When a vehicle is ready to depart, its preceding and following headways are estimated. If the following headway is longer, the vehicle is held by half the difference between the two headways, with the intent of making the vehicle depart when the preceding and following headways are equal. Strategies like this one have been implemented before using information about past departure times to obtain headway estimates. Our implementation uses the performance model to predict headways with dynamic information. Specifically, the model forecasts departure times of the following vehicle from the control vehicle's current stop $d_{v+1,s}$ (to obtain following headway), and of the control vehicle from the preceding vehicle's most recently departed stop $d_{v,e_{v-1}}$ (to obtain preceding headway), as given by

$$h_{v,s} = \max\left(0, \frac{d_{v+1,s} - d'_{v+1,e_{v+1}} - d_{v,e_{v-1}} + t_0}{2}\right) \tag{2.32}$$

OS (Optimization with Static inputs) Hold at selected control points according to the results of the rolling horizon optimization model, with running times and demand inputs defined as period-specific time-independent constants, similar to Delgado et al. (2012). Three periods are defined: one covering the time during which dynamics are in effect and two covering times before and after, when no dynamics are in effect.

OD (Optimization with Dynamic inputs) Hold at selected control points according to the results of the rolling horizon optimization model, with running times and demand inputs defined as time-dependent functions.

2.4 Results and Discussion

The detailed output of the simulation model includes vehicle trajectories and arrival time at the origin, boarding time, and arrival time at the destination for each passenger. From this we can derive performance measures such as waiting time, in-vehicle time, trip time, vehicle loads, holding times, headways, and number of passengers at stops. A probability density function is obtained for each of these performance measures, from which statistics such as mean, standard deviation, and percentiles can be calculated. The principal performance measure of the following analysis is passenger cost.

The trips of passengers who arrive at their origin stop in the analysis period are used to calculate performance measures. Passenger cost and excess waiting time (waiting time in excess of half the scheduled headway) are calculated for each passenger, and the mean cost and excess waiting time across passengers are obtained for each replication. 60 replications were run for each experiment. These results are discussed in Section 2.4.1. Analysis of headways is presented in Section 2.4.2, and analysis of stop crowding in Section 2.4.3. Section 2.4.4 presents an analysis of the sensitivity of performance improvements to stochasticity. Section 2.4.5 explores the value of modeling dynamics when the current state is known perfectly. Section 2.4.6 presents a comparison of computation times for the static and dynamic optimization-based control strategies.

2.4.1 Passenger Cost

Figure 2-3 shows a box-and-whisker plot of mean passenger cost. On the basis of mean passenger cost across replications, EH is consistently better than TH, OS is either better than or similar to EH, and OD is consistently better than or similar to OS.

In the case of dynamic running times and demand and low crowding, small decreases in mean passenger cost are observed going from EH to OS and from OS to OD, though the 90th percentile cost decreases by 0.43 weighted minutes going from EH to OS. In case of dynamic running times and demand and high crowding, passenger cost distributions are similar for the EH and OS strategies, but a significant cost decrease (1.71 weighted minutes or 6.4%) is observed going from OS to OD.

In the case of dynamic running times, static demand, and low crowding, mean passenger costs are similar for the EH and OS strategies, and a very small cost decrease (0.21 weighted minutes or 1.4%) is seen going from OS to OD. Although the maximum cost is greater with OD than OS, the 90th percentile cost is 0.34 weighted minutes less with OD than OS. In the case of dynamic running times, static demand, and high crowding, mean passenger cost decreases by 1.12 weighted minutes (6.3%) going from EH to OS, and by 0.51 weighted minutes (3.1%) going from OS to OD. The 90th percentile cost decreases by 0.46 minutes going from OS to OD. These results suggest that modest performance improvements are achieved by the OD strategy when running times are dynamic, even if passengers are not being left behind due to overcrowding.

In the case of static running times, dynamic demand, and low crowding, a small decrease in mean passenger cost is observed going from EH to OS, and there is no significant difference between the OS and OD cost distributions. This suggests that when dynamics are present only in demand and in the absence of overcrowding, there is no significant benefit to optimizing holding times with dynamic running times and demand. This result is consistent with the finding of Eberlein et al. (2001) that holding policies are not highly sensitive to demand patterns; our results suggest that they are also insensitive to changes in demand over time, unless vehicle capacity is exceeded.

In the case of static running times, dynamic demand, and high crowding, the EH and OS cost distributions are similar, but mean cost decreases by 1.34 weighted minutes (7.2%) and the 90th percentile cost decreases by 1.51 minutes going from OS to OD.

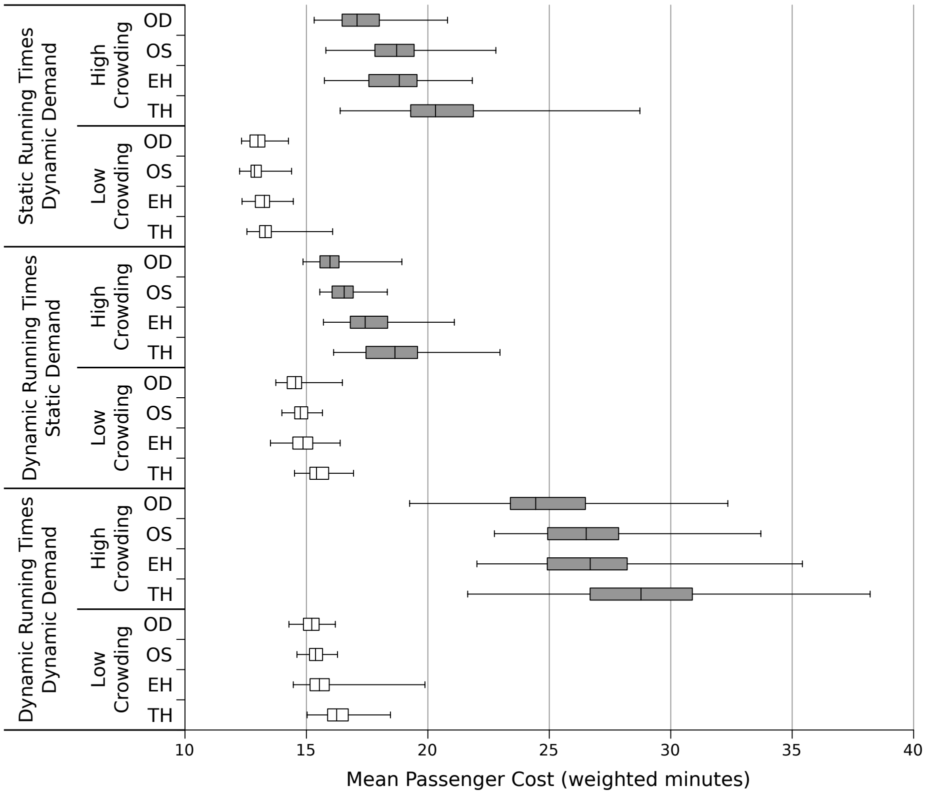


Figure 2-3: Distributions of Mean Passenger Cost

Large performance improvements are seen only in the cases involving dynamic demand and high crowding, which are the two cases having the highest number of passengers left behind. These cases have 6.4% and 7.2% decreases in mean passenger cost going from OS to OD. This suggests that the greatest benefits of optimizing with dynamic running times and demand can be expected when, due to dynamics in running time or demand, a significant fraction of passengers have to wait for more than one vehicle. This is due to the dynamic model's ability to predict when vehicles will be loaded to capacity, a critical factor in determining optimal holding policies, since it is not effective to hold vehicles that will fill up. The OS strategy is equally able to predict when capacities are reached if running times or demand do not vary significantly over the rolling optimization horizon.

Observed differences in excess waiting time closely parallel those in mean passenger cost. Decreases in excess waiting time of 15.8% and 25.0% going from OS to OD are observed for the cases involving dynamic demand and high crowding.

2.4.2 Headways

Table 2.2 shows the mean and standard deviation of headways at the first inbound stop. In all cases, the TH strategy yields the greatest mean headways. Noting that this strategy also yields the greatest mean passenger cost, we observe that performance improvements are possible with more sophisticated holding strategies. The time vehicles spend holding is not productive, and greater overall holding times lead to less frequent service and greater passenger waiting times. However, not holding enough may lead to bunching that is difficult to recover from, which is another way in which productivity can suffer. While the performance advantage of EH over TH stems at least in part from running more frequent service, mean headways increase or remain similar when going from EH to OS and OS to OD in all cases. This suggests that the advantage of OD over OS and OS over EH is not due to running more frequent service.

One might expect that holding optimization necessarily involves a trade-off between frequency of service and headway regularity, since more holding allows decreasing headway variability. The results in Table 2.2 show this is not always the case. For instance, in the case of dynamic running times and demand and low crowding, the OD strategy reduces both the mean and the standard deviation of headways with respect to the TH strategy. This indicates that some strategies choose holding times more effectively.

One might also expect that for a given mean headway, lower headway variabilities lead to better performance. This notion, captured formally in the commonly used equation expressing expected waiting time in terms of headway mean and standard deviation, assumes that passengers are not left behind by (full) vehicles and that headway variability is entirely random (rather than systematically time-dependent). In the two cases with the greatest decreases

Table 2.2: Headways at First Stop

Running Times	Demand	Crowding	Strategy	Mean	Standard Deviation
Dynamic	Dynamic	Low	TH	5.16	2.52
			EH	4.79	2.46
			OS	5.04	1.96
			OD	4.93	2.07
Dynamic	Dynamic	High	TH	5.73	2.78
			EH	5.46	2.67
			OS	5.47	2.74
			OD	5.43	2.97
Dynamic	Static	Low	TH	5.11	2.30
			EH	4.76	2.31
			OS	4.95	1.98
			OD	4.90	1.94
Dynamic	Static	High	TH	5.78	2.45
			EH	5.44	2.58
			OS	5.69	1.68
			OD	5.54	1.51
Static	Dynamic	Low	TH	4.52	1.71
			EH	4.28	2.04
			OS	4.42	1.59
			OD	4.40	1.71
Static	Dynamic	High	TH	5.13	1.43
			EH	4.83	1.91
			OS	5.02	1.55
			OD	4.97	2.11

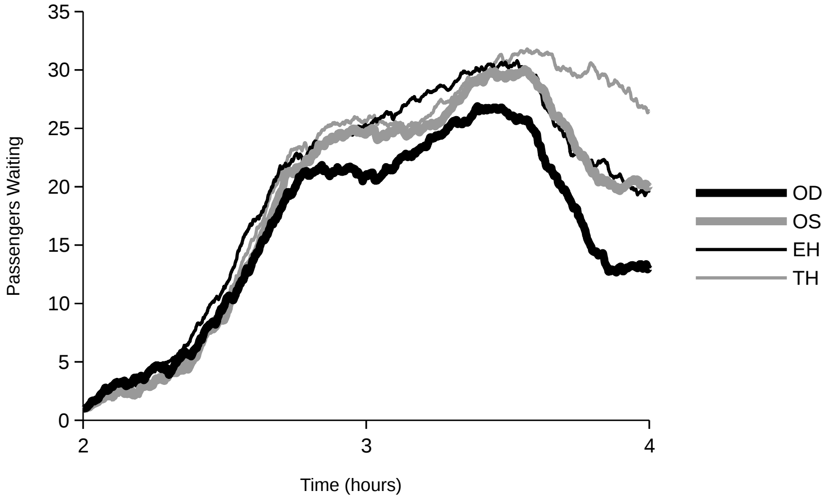


Figure 2-4: Stop Crowding at Maximum Load Point

in mean passenger cost going from OS to OD, which are also the two cases with dynamic demand and high crowding, mean headways are similar but headway variability increases going from OS to OD. In these cases a significant number of passengers are left behind, so the best holding policies are not necessarily those that balance headways. Indeed, there may be situations in which strategic bunching is desirable for a period of time to manage overcrowding.

2.4.3 Crowding at Stops

The number of passengers waiting at the maximum load point (stop 10 in the first direction) was analyzed in the case of dynamic running times and demand and high crowding. Figure 2-4 shows mean loads over time for the different control strategies.

The OD strategy leads to the lowest stop crowding throughout most of the analysis period, ending at about 13 passengers compared to about 20 passengers for the EH and OS strategies. The difference in number of passengers at this time is as large between OS and OD as between TH and EH. This suggests that the OD strategy is more effective in controlling overcrowding at stops in services running at capacity with significant running time and demand dynamics.

2.4.4 Sensitivity to Stochasticity

The earlier discussion shows that the OD strategy can lead to improved performance in some cases. There are at least two possible reasons for this. First, since the OD strategy can predict future system states more accurately, it may

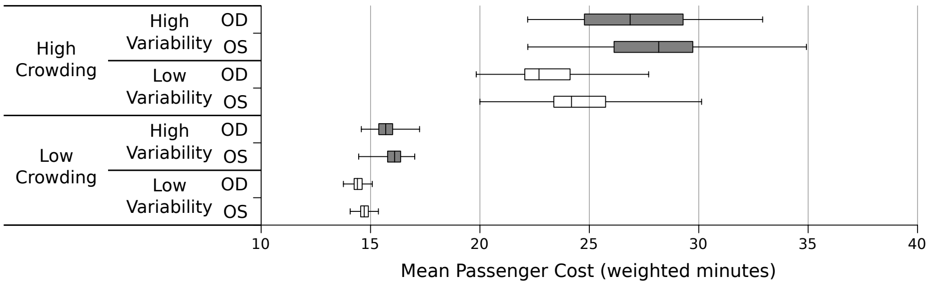


Figure 2-5: Distributions of Mean Passenger Cost for Low and High Running Time Variability

be able to generate preemptive control policies that prepare the system for delays or surges in demand before these reach a critical level. Second, although the forecasting model underlying the OS and OD strategies is deterministic, the OD strategy may be able to react to disruptions caused by stochastic running times and demand more effectively than the OS strategy, by considering dynamics of running times and demand. In order to understand which of these plays the greater role, we conduct a sensitivity analysis of performance using the OS and OD holding strategies under low and high running time variability. The analysis focuses on the two cases having dynamic running times and demand (with low and high crowding).

In all previous cases, running times between consecutive stops are drawn from a log-normal distribution with a mean of 60 seconds and a coefficient of variation of 0.4 (shifted for dynamics). We introduce low and high variability scenarios that instead have coefficients of variation of 0.2 and 0.6, respectively. In all cases, demand follows the same Poisson process as before. Figure 2-5 shows mean passenger cost distributions in a box and whisker plot.

Both mean passenger costs and variability in these means across replications increase with running time variability. As before, the OD strategy yields lower mean passenger costs than the OS strategy, and the improvement is small in cases with low crowding and significant in cases with high crowding. Going from OS to OD, mean passenger costs decrease by 0.28 weighted minutes in the low crowding, low variability case, 0.33 weighted minutes in the low crowding, high variability case, 1.31 weighted minutes in the high crowding, low variability case, and 1.20 minutes in the high crowding, high variability case. For low crowding, the improvement is 0.05 weighted minutes greater in the high variability case than the low variability case. For high crowding, the improvement is 0.11 weighted minutes less in the high variability case than the low variability case. The small magnitude of these two differences suggests that performance improvements going from OS to OD are robust to changes in system stochasticity, and that most of the improvements come not from the OD strategy reacting with greater aptitude to unpredictable stochasticity, but from its ability to generate preemptive policies in light of forecast dynamics in running times and demand.

2.4.5 Sensitivity to Current State Information Accuracy

Since the OD strategy incorporates time-varying running times and demand, it is better able to estimate the current system state as well as predict future states under a variety of control policy scenarios. The model used for holding time optimization takes inputs that specify both the current system state and the running time and demand functions that govern how the system is expected to evolve. For instance, the number of passengers initially waiting at stops (ι'_{s_b,s_a}) or inside vehicles (l'_{v,s_a}), by origin-destination pair, are given to the optimization model. The OD strategy can estimate these quantities better because it has more accurate descriptions of passenger arrival rates over time.

We have seen that the OD strategy improves performance in some cases, but the question remains: to what extent do performance benefits stem from the ability to estimate *current* system state more accurately rather than from the ability to predict *future* system states with dynamic running times and demand? If the benefit came mostly from the former, then performance under the OS strategy could be improved with the use of passenger counting technologies. To investigate, we evaluate operations in the case of dynamic running times and demand and high crowding using the OS and OD strategies, but replacing the estimation of number of passengers at stops and in vehicles by the true values known to the simulation model. In reality it would be difficult to obtain this information, but the result of this experiment allows us to understand the relative importance of accuracy of current vs. future system states.

Figure 2-6 shows mean passenger cost distributions for the case of dynamic running times and demand and high crowding in three different scenarios having different information about the current state. In the first case, the number of passengers is estimated both in vehicles and at stops. This is the base case presented in Section 2.4.1, repeated here for ease of comparison. In the second case, the number of passengers in vehicles (by destination) is known perfectly but the number of passengers at stops (by destination) is estimated. In the third case, the number of passengers, both in vehicles and at stops, is known perfectly.

The improvement going from the OS strategy to the OD strategy is 1.71 weighted minutes in the base case, 1.31 weighted minutes in the case of known vehicle loads but estimated number of passengers at stops, and 0.74 weighted minutes in the case of known vehicle loads and number of passengers at stops. As expected, the improvement decreases as more current state information is known rather than estimated, because the information accuracy gap closes. The 0.40 weighted minutes of improvement gained by knowing the number of passengers initially in vehicles is 24% of the 1.71 weighted minutes improve-

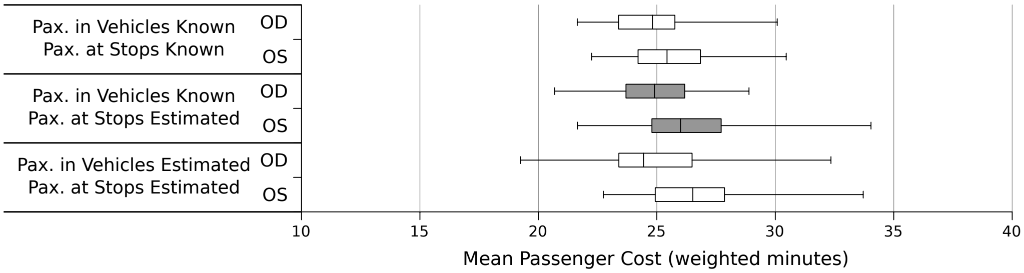


Figure 2-6: Distributions of Mean Passenger Cost with Perfect Current State Information

ment in the base case. The 0.97 weighted minutes improvement gained by knowing the number of passengers initially in vehicles and at stops is 57% of the 1.71 weighted minutes improvement in the base case. In other words, more than half of the OD strategy's advantage appears to come from estimating the current state, rather than future states, more accurately.

Focusing on the scenarios with the OD control strategy, there is negligible performance difference between the base case and the case of known vehicle loads but estimated number of passengers at stops, and a very small improvement of 0.13 weighted minutes going from that to the case of known number of passengers in vehicles and at stops. This indicates that the OD strategy's current state demand estimates are sufficiently accurate for this application.

2.4.6 Computation Time

In the context of real-time control, it is important to consider not only the effectiveness of control but also the computational effort required to generate control policies. This is because in real applications decisions must be made quickly and provided to vehicles soon after their arrival at control points. Simulations were run on a computer having an Intel Core i7-3930K processor running at 3.20GHz. Mean computation times were 0.77 seconds with OS and 1.08 seconds with OD. The 90th percentile computation time for the OD strategy was 1.83 seconds, and the maximum was 3.42 seconds. Therefore, it is feasible to employ the OD strategy in real-time applications.

2.5 Concluding Remarks

The mathematical model formulated in this research captures dynamic running times and demand. A holding strategy based on this model (OD) was tested in a simulation environment under a variety of cases having different dynamics and its performance was compared to that of three other control strategies, including a similar optimization-based strategy that assumes the equivalent static running times and demand (OS). The principal performance measure used for evaluating strategies was mean passenger cost, which reflects passenger trip durations, weighting time spent waiting at stops twice as much as time spent in the vehicle. Excess waiting time, headways, vehicle loads, and crowding at stops were also considered. The key findings are:

1. The proposed control strategy based on optimization with dynamic inputs (OD) outperforms the OS strategy in cases where, due to running time or demand dynamics, the system becomes overcrowded and passengers are left behind by (full) vehicles. In cases having dynamic running times but not significant overcrowding, the proposed dynamic strategy OD modestly outperforms the OS strategy. Dynamics in demand do not provide an opportunity for the OD strategy to improve performance over the OS strategy, unless the dynamics lead to significant overcrowding.

2. Optimization-based control strategies lead to similar, or better, performance than the even headway strategy (EH), and the EH strategy outperforms the (fixed) target headway strategy (TH).

3. Holding strategies improve performance through a combination of running more frequent service due to more efficient use of the fleet and regulating headways. The performance improvement going from TH to EH is at least in part due to running more frequent service. However, headway regulation is the principal mechanism by which performance improves going from EH to OS and from OS to OD.

4. When dynamics are present, it is possible for a holding strategy to improve performance even if its holding policies lead to increased headway variability (and similar mean headways), especially if the increase in headway variability is a byproduct of more efficient use of resources and the optimal holding policy is different from merely balancing headways. This is relevant in cases of overcrowded operations.

5. The degree to which the OD strategy outperforms the OS strategy is not sensitive to running time variability. Most of the benefit comes from the OD strategy's ability to generate preemptive holding policies in light of forecast dynamics.

6. A large part of the performance improvement going from OS to OD is due to more accurate estimates of the current state, while the remainder comes from modeling running times and demand dynamically when forecasting future system states under varying holding policy scenarios.

7. Computation times with the OD strategy are suitable for real-time application.

The findings presented in this chapter are specific to the hypothetical simulated transit service, although they might also hold for many real high-frequency transit services. Future research could explore the effectiveness of the proposed dynamic holding strategy in a real transit service. In real applications it may be undesirable to delay vehicles whose drivers are expected to arrive late at their scheduled relief point. The incorporation of crew constraints and their effect on control effectiveness could be explored. The effect of intentional or benign disregard of holding instructions by drivers, and of errors on the information assumed on vehicle locations, running times, and passenger arrival rates could be investigated. Control strategies that capture the stochastic aspects of transit performance could be developed and tested, and the relative importance of capturing stochasticity vs. capturing dynamics could be investigated.

Chapter 3

Holding Control Under Event Driven Dynamics

Operating high-frequency transit without controlling departure times from terminals and stops can lead to disrupted service, often in the form of headway variability and bunching, which in turn leads to passengers experiencing long waiting times, crowded vehicles, and unreliable trip times (Delgado et al., 2012). A number of factors can disrupt service, including variability in running times and demand, inadequate operations planning or execution, and atypical events that alter operating conditions. Control is exercised in response to disrupted service, typically by observing current operating conditions through automated data collection systems and generating a control response based on heuristics or optimization models.

This chapter focuses on holding control responses to disruptions triggered by events. Events such as road congestion caused by traffic accidents, rail signal failures, and medical emergencies in vehicles or stations are unpredictable. Other events such as concerts and sport contests may cause predictable surges in demand, resulting in short-term local congestion. Regardless, information about the event can be considered in the generation of a control response. Although events such as traffic accidents cannot be predicted, the detection of the event along with an estimate of its duration based on past experience and real-time updates on the progress towards resolution can be used to predict congestion and how it might affect transit service. The end of a concert or sports event is associated with an increase in traffic and demand for transit service. The time at which the event ends, associated traffic delays, and number of people who will take transit can be estimated with information about event attendance, real-time updates on the progress of the event, and observations of operations during similar events in the past.

These predictions can be used to apply control preemptively instead of waiting until service is disrupted enough that a problem is apparent. Advances in sensing and telecommunications technology have fostered an increasingly data-rich environment in which data about many aspects of a transit system can be collected in real time from multiple sources such as infrastructure health monitoring systems, traffic sensors, live camera feeds, automated vehicle location, automated fare collection, automated passenger counting, and crowd-sourcing through social media platforms. However, the potential value of information derivable from these data sources to improving the effectiveness of real-time control responses to events has not been explored.

An additional factor enabling control with information about events is the holding control optimization model presented in Chapter 2, which captures expected dynamics in running times and demand. This research applies that model in the context of dynamics driven by unforeseen and foreseen events, with the objective of evaluating the potential value of information about event dynamics on real-time control effectiveness.

It is not possible to respond to unforeseen events through traditional operations planning, because it happens before service delivery. For large foreseen events, transit operators may change the operations plan so that, for example, extra vehicles are operating during a large event. Dealing with these events through real-time control is no substitute for good operations planning. However, once extra resources are deployed, it is necessary to control them to ensure that they are used most effectively. Some predictable events are not large enough to warrant a special operations plan, but can still affect service in the absence of real-time control. This research aims to show to what extent and in what manner real-time information about events can improve the effectiveness of real-time control in both these contexts.

This chapter is organized as follows: Section 3.1 reviews the literature, Section 3.2 presents the framework, Section 3.3 presents the holding time optimization model, Section 3.4 discusses the application of this model to a simulated transit service subjected to both unforeseen and foreseen events, and Section 3.5 draws conclusions.

3.1 Literature Review

The majority of control strategies in the literature do not utilize information about events. Strategies that do not consider the availability of real-time information (Osuna and Newell, 1972, Barnett, 1974, Newell, 1974, Turnquist and Blume, 1980, and Abkowitz and Lepofsky, 1990) do not capture event-driven dynamics. Strategies aimed at schedule adherence (Adamski and Turnau, 1998) are only suitable for long headway services in which passengers are aware of schedules and time their arrival at stops to take specific vehicle trips. Strategies aimed at headway adherence (Rossetti and Turitto, 1998) or headway regularity (Daganzo, 2009, Daganzo and Pilachowski, 2011, Cats et al., 2011, and Bartholdi and Eisenstein, 2012) are sensitive only to local current conditions such as the preceding and following headways of a vehicle, so they do not reflect event-driven dynamics.

Strategies based on cost minimization (O'Dell and Wilson, 1999, Eberlein et al., 2001, Shen and Wilson, 2001, Zhao et al., 2003, Sun and Hickman, 2008, Puong and Wilson, 2008, Delgado et al., 2009, Delgado et al., 2012, Sáez et al., 2012, and Chapter 2) employ mathematical programming or other optimization methods and a variety of information including vehicle locations, loads, and passenger arrival rates, sometimes in a rolling horizon formulation. They respond to both the current state and predicted future states under varying control scenarios. Most assume static running times and demand, making it cumbersome, if not impossible, to capture event-driven dynamics and preemptively consider them in the production of a control plan. Others (O'Dell and Wilson, 1999, Shen and Wilson, 2001, and Puong and Wilson, 2008) formulate holding and short-turning optimization models and apply them to temporary blockages in rail lines with static running times and demand.

Within the family of real-time control strategies based on future system state prediction models, the strategy presented in Chapter 2 models running times and demand dynamically. The effectiveness of the model is evaluated by controlling a simulated transit service exhibiting time-dependent running times and demand. The results show that the dynamic control strategy outperforms its static equivalent in high-demand cases where passengers may be left behind at stops, and also when running times change significantly over time. However, the model is only evaluated in scenarios having dynamics that are known well in advance, such as those caused by the typical rush hour: running times and demand rising and falling gradually.

This chapter applies the model developed in Chapter 2 to scenarios in which dynamics are driven by events including both unexpected incidents and anticipated surges in demand. The following sections present the dynamic model framework and an evaluation of its performance under event-driven dynamics.

3.2 Framework

Controlling transit operations considering event-driven dynamics can potentially improve the effectiveness of control, reducing waiting times and trip times for passengers during an event. The performance benefit of a strategy that captures event-driven dynamics, i.e. an *informed* strategy, is derived from its *information advantage* over a strategy that neglects event-driven dynamics, i.e. a *naive* strategy. Having more realistic predictions of running times and demand can lead to better predictions of future system states (and their cost to passengers) under different control scenarios, ultimately leading to more effective control policies. In order to apply holding control considering the effect of events, transit operators must (1) be aware of events, (2) gather relevant data, (3) model future operating conditions, and (4) optimize holding times capturing event-driven dynamics.

Handling unforeseen events requires detecting the event in real-time. Since there is little time to act, operators must rely on data that can be obtained quickly; it may be difficult to obtain details. For example, an operations control center might detect a signal failure on a rail transit link, but know only the time and location of the event. Handling planned events requires knowing when the event is scheduled to occur and identifying potential sources of data that help inform how the event might affect operations. The data, which may be specific to the event, can be gathered in advance. For example, relevant data for a concert might include date and time, estimated attendance, and estimated duration.

Once relevant data is available, the operator must estimate and model future operations in light of the event. Previous experience with similar kinds of events can help. Historical data on running times and demand during events can be combined with event-specific data to prepare predictive models. Care should be taken to separate exogenous factors from those triggered by control responses. For example, running time models should be based on running time observations excluding holding times. Later, when a new event of the same kind occurs, these models can be used to estimate new conditions. For example, a simple model might predict delays caused by a traffic accident according to its reported severity, basing its predictions on running times observed during similar events in the past. A different model could be used to predict the demand surge at the end of a concert based on estimated attendance and historical data pairing attendance with observed demand surges.

Unforeseen events must be handled quickly, because the information advantage of an informed control strategy can be ephemeral. Naive control strategies can respond effectively to disruptions already visible in the current state of a transit line. The advantage of informed control is predicting disruption effects before they materialize, which can happen quickly after an event is detected. Since the information advantage is determined not only by more realistic predictions of future system states but also by how far in advance this knowledge is utilized, the advantage of informed control for unforeseen events can be significantly less than for foreseen events. In order to maximize the information advantage, it would be advantageous to automate the process of detecting unforeseen events, estimating changes in operating conditions, and updating the control optimization model.

Events cause transients in running times and demand, which is why it is important to model these changes dynamically. Running times and demand are specified as functions of time, allowing the optimization model to assign numbers of boarding passengers and running times between stops to vehicles according to their departure times from stops. General piecewise functions can be used, making the framework flexible enough to handle a wide variety of events, including traffic, blocked links, demand surges, etc. The functions can be updated as new information becomes available. For example, at the beginning of a signal failure the transit operator might only know its location, so the running time function for the affected link can be changed to reflect a blockage lasting some estimated amount of time. Once the operator identifies the cause of the signal failure and the availability of crew required to fix the signal, the duration estimate can be refined and the running time function updated accordingly. This allows control optimization to utilize information as it becomes available.

3.3 Control Model

The dynamic model presented in Chapter 2 is well-suited to capture transients in running times and demand caused by events, in contrast to static models that assume constant running times and demand throughout the prediction horizon. The objective is to minimize passenger cost over the prediction horizon, which combines waiting time and in-vehicle delay, subject to constraints on holding times at stops. For this research, the objective function is modified to minimize total, rather than mean, passenger cost, by not dividing the sum of waiting and in-vehicle time by the number of passengers boarding during the prediction horizon. Minimizing mean cost is necessary in rolling horizons defined temporally in order to discourage holding policies that slow vehicles down to reduce total cost by serving fewer passengers. Since in this research the rolling horizon is defined spatially, the horizon includes a full cycle for each vehicle regardless of how long it takes. Therefore, minimizing total cost encourages running fast, instead of slow, to serve fewer passengers. Minimizing mean cost instead can encourage excessive holding when vehicles do not run full, because by slowing down service, more passengers are served in the prediction horizon. This is not the case in overcrowded lines. As before, the control model is used to optimize holding times, for all vehicles at all stops over the rolling prediction horizon, every time a vehicle arrives at a control point. The holding time suggested by the optimizer for the vehicle that triggered the modeling-optimization event at its current stop is implemented.

3.4 Applications

This section presents the application of the framework and control model to a simulated high-frequency transit service subject to changes in the operating environment caused by events. Two cases are explored: one in which the event is unforeseen and information about it becomes available only at the time of the event, and another in which the event is foreseen and information about it is known well in advance. For each case, two control strategies are applied: a *naive* strategy that ignores information about the event and an *informed* strategy that utilizes it. The naive strategy can detect and respond to disruptions as they develop, but without awareness of the event, it assumes typical running times and demand to forecast future system states. The informed model anticipates changes in the operating environment and controls preemptively. The value of information about events is assessed by comparing transit service performance under the naive and informed control strategies.

The transit line used in both applications is a simple bus route having 20 stops per direction. Running times between stops are log-normally distributed with a mean of 1 minute and a coefficient of variation of 0.3. Demand is modeled using a non-stationary Poisson process, with all origin-destination pairs in each direction having the same arrival rate. Vehicles can carry up to 60 passengers. Holding is allowed at stops 5, 10, 15, and 20 in each direction. Holding at the first three control points in each direction is limited to 2 minutes. 100 replications are used for each simulation.

The simulation model outputs vehicle trajectories and arrival time at the origin, boarding time, and arrival time at the destination for each passenger. From this we can derive performance measures such as waiting time, in-vehicle time, trip time, vehicle loads, headways, and number of passengers at stops over time. A probability density function is obtained for each of these performance measures, from which statistics such as mean, standard deviation, and percentiles can be calculated. Performance is evaluated based mainly on passenger cost (twice the waiting time plus in-vehicle time).

3.4.1 Unforeseen Event

The first application deals with an unforeseen event causing a link to be blocked for a short time during rush hour. This might be due to a traffic accident. Traffic quickly builds up, causing delays on upstream links. Passengers continue to arrive as usual, but they have longer waiting times due to the long headway caused by the event. Since the event cannot be anticipated, no information about the incident is known before it occurs. However, once the incident occurs, it is reported and information about it becomes available. For this application the information of interest is the duration of the incident. We optimistically assume that the informed model becomes aware about the incident and the duration of the blockage at the time the incident occurs. In real applications there might be a delay before awareness and an error in the estimated duration of the blockage, though the estimate could be updated based on incident monitoring.

The case, illustrated in Figure 3-1, is modeled by adding a deterministic dynamic running time delay. The delay first occurs on the link connecting stops 4 and 5 in the first direction, but then it propagates to the three links upstream at 1 minute offsets to simulate traffic congestion building up. The delay is

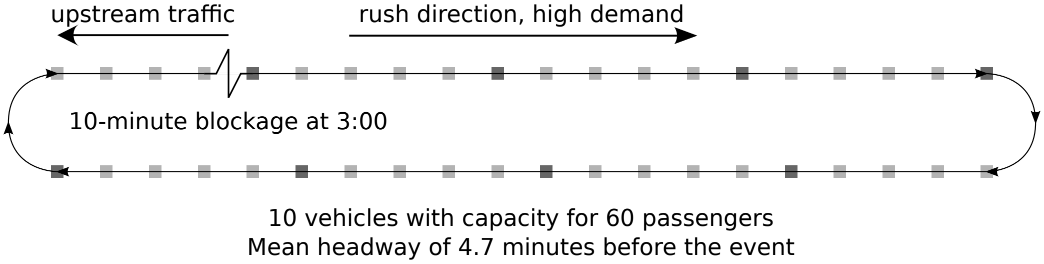

upstream traffic
rush direction, high demand
10-minute blockage at 3:00
10 vehicles with capacity for 60 passengers
Mean headway of 4.7 minutes before the event

Figure 3-1: Unforeseen Event Case

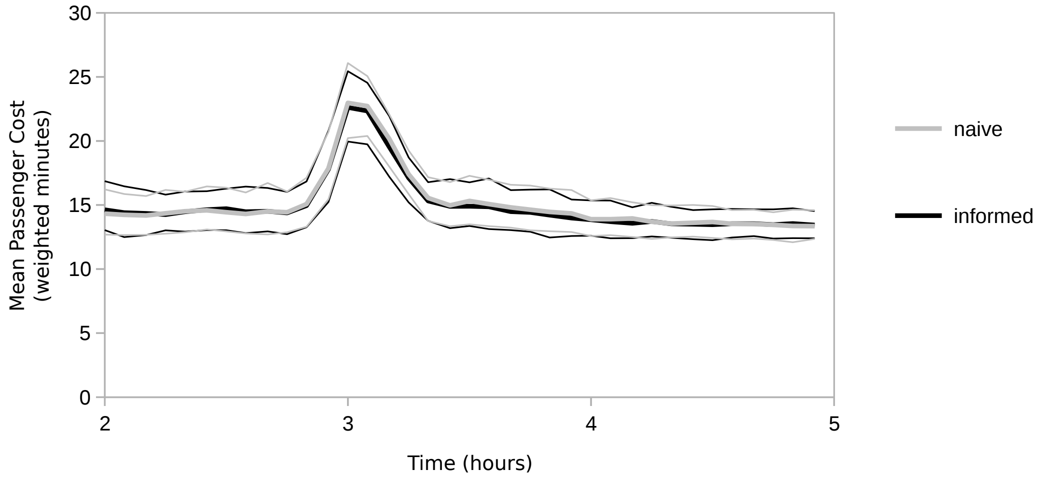


Figure 3-2: Mean Passenger Cost for the Unforeseen Event

initially 10 minutes but tapers down to 0 over a period of 10 minutes, such that a vehicle entering the link 5 minutes after the blockage begins is delayed for only the 5 remaining minutes. Demand is higher in the first (rush) direction than the second, with vehicles reaching about 85% of capacity on average before the event. Demand in the first direction begins decreasing shortly after the event, simulating the transition from peak to mid-day operations. The line operates with 10 vehicles. The mean headway before the event is 4.7 minutes.

Neither of the control models is aware of the event before 3:00. When the blockage begins (at 3:00), the informed model becomes aware of the event-driven delays in running times, but the naive model continues using the typical running times. Hence, the informed model has the advantage of being able to predict the disruptions caused by the link blockage before the disruptions can be detected in the current state. The advantage lasts only for a few minutes, because once vehicles are delayed, the naive model sees the effect of the blockage.

Mean cost for all passengers drops from 14.6 to 14.5 weighted minutes (less than a 1% reduction) going from the naive strategy to the informed strategy. This difference is small but statistically significant at a significance level of 0.01. Since the system is stochastic, outcomes vary across replications. Figure 3-2 shows the mean, 10th percentile, and 90th percentile of the probability distribution of mean passenger cost over time for the naive and informed control models, considering passengers boarding at all stops. Thick lines show the across-replications rolling mean, while thin lines show the across-replications 10th and 90th rolling percentiles. The plotted statistics are computed by dividing the analysis period beginning at 2:00 and ending at 5:00 into 36 5-minute intervals, then calculating the mean cost across passengers arriving at their origin stop in each interval for each replication, and finally computing the mean, 10th percentile, and 90th percentile of that across replications. The differences in cost achieved by the two control models are small at all times.

Since Figure 3-2 groups passengers by the time they arrive at their origin stop, it mixes passengers affected by the blockage with others who are not. The timing of affected passengers is later the farther downstream from the blockage their origin stop is. Accounting for the required timing offsets, passengers directly affected by the blockage enjoy 0.3 minutes of waiting time savings under the informed strategy (a 4% reduction).

3.4.2 Foreseen Event

The second application deals with a planned event that induces running time delays and a surge in demand with predictable timing and magnitude. The event might be, for instance, the end of a concert or sport contest, when those attending the event leave the venue and crowd the surroundings. Traffic slows near the venue due to automobiles leaving and the great number of people crossing streets. A portion of the attendees take the transit service analyzed here. Some have to wait for more than one vehicle because vehicles reach capacity.

The case, illustrated in Figure 3-3, is modeled by adding a surge of demand at the midpoint of the route in both directions, in addition to a running time delay in the adjacent links. The demand surge is modeled by temporarily increasing the arrival rate governing the Poisson process for passengers originating at stop 10 in each direction. The surge accounts for two full buses per direction over 15 minutes, starting suddenly at 3:00 and tapering off linearly back to the base arrival rate. A deterministic running time delay of 3 minutes over the same period is added to the two links arriving at and departing from stop 10 in each direction. The service operates with 5 vehicles. The mean headway before the event is 7.8 minutes.

The informed model is aware of the demand surge and running time delays even before the event. The information enters the model gradually as the prediction horizon end rolls past the event time. In contrast, the naive model assumes typical operating conditions and only responds after disruptions are evident in the current state.

Space-time diagrams of naive and informed control showing vehicle movement and loads (by line thickness) for a single replication are shown in Figure 3-4. Stops in both directions are shown on the vertical axis and time

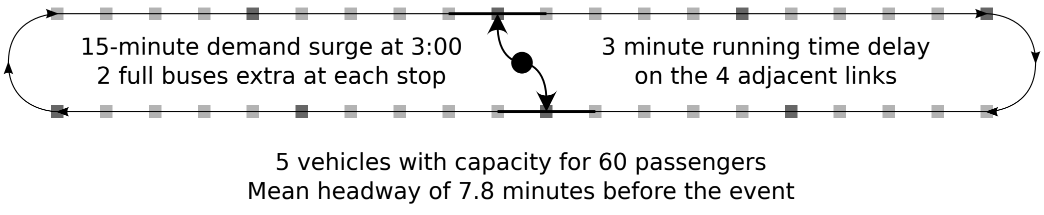
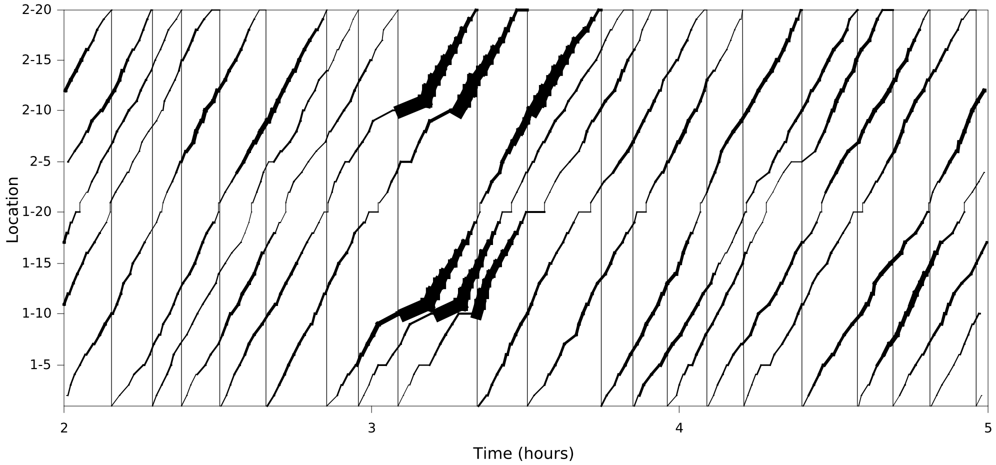
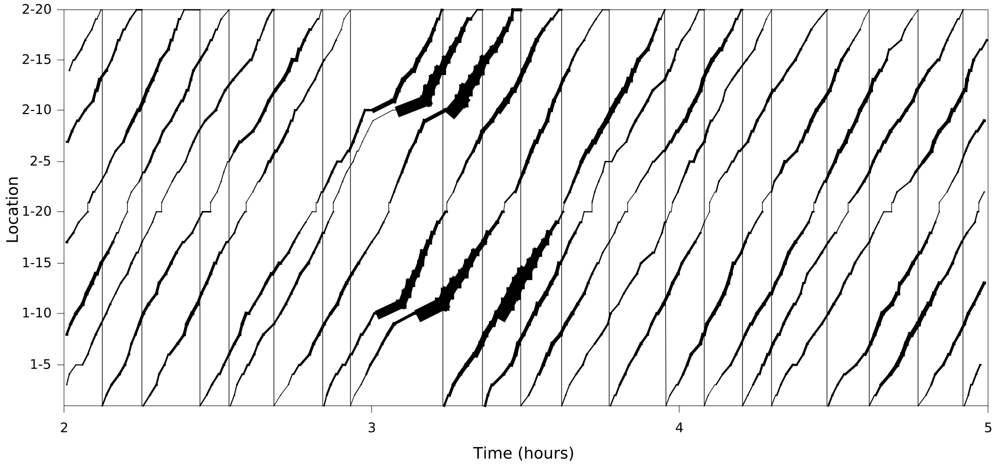

15-minute demand surge at 3:00
2 full buses extra at each stop
3 minute running time delay
on the 4 adjacent links
5 vehicles with capacity for 60 passengers
Mean headway of 7.8 minutes before the event

Figure 3-3: Foreseen Event Case



(a) Naive Control



(b) Informed Control

Figure 3-4: Vehicle Trajectories for the Foreseen Event

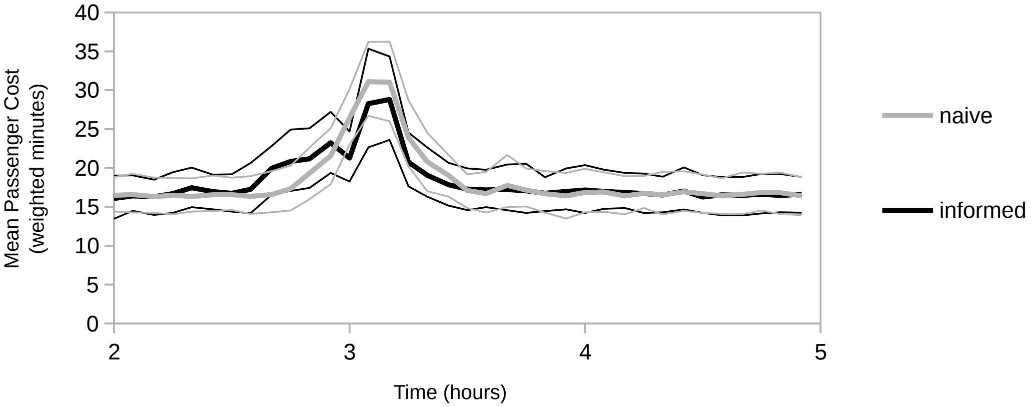


Figure 3-5: Mean Passenger Cost for the Foreseen Event

is shown on the horizontal axis. The naive model does not react until after vehicles are noticeably delayed, so a long headway results between the first delayed vehicle and the preceding one. Holding is applied after the event to regulate headways. The informed model begins responding preemptively, so that by sacrificing some headway regularity before the event, vehicles arrive shortly after the event, reducing waiting time. In the illustrated example (one of the 100 replications), the vehicle that departs stop 10 in the second direction right after 3:00 is held more than what is required to regulate headways almost a full cycle before, at stop 20 in the second direction, and the following vehicle is not held at all, although some holding would have been necessary to achieve regular headways in the first direction. Consequently, vehicles arrive soon after the demand surge begins with shorter headways to serve passengers from the event in both directions, and headways are regular after the event.

Figure 3-5 shows the mean, 10th percentile, and 90th percentile of the probability distribution of mean passenger cost over time for the naive and informed control models, considering passengers boarding at all stops. Thick lines show the across-replications rolling mean, while thin lines show the across-replications rolling 10th and 90th percentiles. Informed control leads to higher mean cost shortly before the event (when there are few passengers in the vehicles held) because of preemptive holding, but lower mean cost over the half hour following the event, when many more passengers are affected. Costs are very similar both well in advance of the event and starting a half hour following the event. Mean cost for passengers boarding at the two stops affected by the event during the event-induced demand surge drop from 30.2 to 24.8 weighted minutes going from naive to informed control (an 18% reduction). Mean cost for all passengers boarding anywhere (including passengers affected by holding before the event) drops from 19.8 to 19.0 weighted minutes going from naive to informed control (a 4% reduction).

Figure 3-6 shows the mean, 10th percentile, and 90th percentile of the probability distribution of number of passengers waiting, over time, at stop 10 in

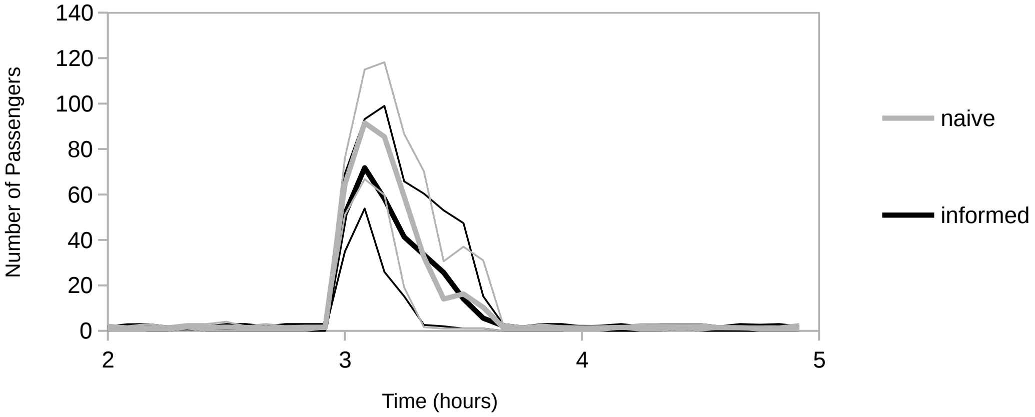


Figure 3-6: Passengers Waiting at Stop 10 for the Foreseen Event

the first direction, for the naive and informed control models. Thick lines show the across-replications mean number of passengers and thin lines show the across-replications 10th and 90th percentiles. The informed control model achieves a lower number of passengers waiting during almost all the demand surge. The mean peak number of passengers is reduced from 91.5 to 71.8 (a 22% reduction). In fact, the mean peak number of passengers with informed control is only 5 passengers more than the 10th percentile number of passengers with naive control. Few passengers are waiting at any given time before the event or after the demand surge dissipates.

Figure 3-7 shows the mean holding intensity over time for the naive and informed control models, considering all passengers. Holding intensity, measured in passenger-minutes per minute, indicates how much passengers are held and increases with both holding duration and number of passengers held. Informed control leads to more holding during the period leading to the demand surge but less holding during the demand surge, when it is desirable to hold less in order to increase supplied capacity. Holding intensity is the

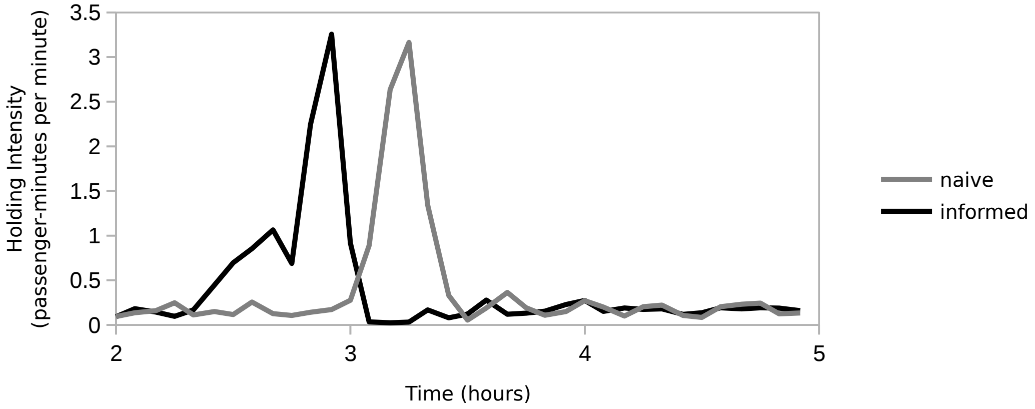


Figure 3-7: Holding Intensity for the Foreseen Event

same under both strategies far before the event as well as after the demand surge dissipates. Mean holding intensity over the entire analysis period is 0.4 passenger-minutes per minute for both strategies, suggesting that the benefits of informed control do not require additional holding.

3.4.3 Effect of Erroneous Information

In the cases presented so far, the informed strategy is based on perfect information about the event. There are no errors in the estimated timing, form, and magnitude of disruptions caused by the event. The results presented so far can be interpreted as upper bounds on the value of information for real-time holding control during events, because errors in the provided information can harm the performance of the informed strategy. In real cases, the effects of an event on running times and demand are uncertain a priori. Focusing on the foreseen event case, this section explores the effect of erroneous information on the effectiveness of the informed control strategy. Two types of errors are considered: errors in the magnitude of the demand surge and errors in its timing. While the information used by the informed control strategy changes, the actual magnitude and timing of the demand surge and associated running time delays remain as before.

To investigate the effect of errors in the estimated magnitude of the demand surge, the optimization model is given dynamic demand functions accounting for either 1 full bus or 3 full buses of additional passengers in the stops affected by the demand surge, instead of the actual 2 full buses of additional passengers used by the simulation model. To investigate the effect of errors in the estimated time of the demand surge, the optimization model is given dynamic demand and running time functions for the event occurring either 10 minutes early or 10 minutes late. With errors in estimated time there is opportunity to correct the information. When the model operates assuming the event will happen 10 minutes early, the information is corrected 10 minutes prior to the actual event. When the model optimizes assuming the event will happen 10 minutes late, the information is corrected at the time of the actual event.

Table 3.1 shows the percentage improvements in mean passenger cost with respect to the cost achieved by the naive strategy, for all passengers as well as for only surge passengers, i.e. those arriving at the two stops affected by the demand surge during the surge. Errors in the magnitude estimate are less onerous than errors in timing. The informed strategy achieves lower cost relative to the naive strategy with under- or overestimated surge magnitude, but the errors in timing increase the cost relative to the naive strategy. This suggests that when there is significant uncertainty in the timing of an event, it is better to ignore the information until there is higher certainty.

Table 3.1: Effect of Erroneous Information for the Foreseen Event

Information Error	% Improvement All Passengers	% Improvement Surge Passengers
None (Perfect Information)	4	18
Lower Demand Surge Magnitude	4	13
Higher Demand Surge Magnitude	1	10
Earlier Demand Surge	−4	−2
Later Demand Surge	−3	−1

3.5 Concluding Remarks

The holding control model presented in Chapter 2, based on rolling-horizon optimization capturing dynamics of running times and demand, can be applied to control a high-frequency transit service subject to dynamics caused by events. Two types of events are considered: unforeseen events such as a link blockage due to a traffic accident, for which dynamics can be predicted only after the event has been detected, and foreseen events such as the end of a concert or sport contest, for which demand and running time dynamics can be predicted in advance.

This research compares the performance of a simulated transit service subject to foreseen and unforeseen events under two control scenarios: naive and informed control. Naive control generates holding policies based on rolling horizon predictions assuming typical conditions, without the dynamics of the event coming into play. Informed control generates holding policies based on rolling horizon predictions capturing predicted event dynamics. Therefore, it can begin responding to events before disruptions develop. The performance improvements achieved by the informed control model can be interpreted as the value of information about event dynamics for real-time control.

Information about events can improve the effectiveness of real-time holding control in response to events. The magnitude of the performance improvement depends on the information advantage of informed control over naive control. Since the advantage grows as information is known farther in advance, performance improves more with foreseen events than with unforeseen events. Preemptive holding suggested by the informed control model increases passenger cost shortly before a foreseen event affects service, but decreases passenger cost thereafter, for an overall net benefit.

Information about an unforeseen link blockage does not significantly improve the effectiveness of control. In the specific case presented, a cost reduction of less than 1% was observed. Information about a foreseen localized demand surge can improve the effectiveness of control by decreasing waiting time of passengers from the event and number of passengers waiting at the affected stops. In the specific case presented, an 18% cost reduction was observed for passengers arriving at the two event stops during the demand surge, and the peak number of passengers waiting at the affected stops decreased by 22%. These improvements were observed across the distribution of outcomes, including low and high percentiles of passenger cost and number of passengers waiting at the stops affected by the demand surge. These results reflect the performance of an informed control strategy given error-free information.

Erroneous information decreases the effectiveness of the informed control strategy. In the case of a foreseen demand surge, errors in the surge time estimate are more onerous than errors in the surge magnitude estimate. Controlling assuming incorrect event timing can be counterproductive. It is better not to use information about an event, i.e. to use the naive strategy, until its timing is certain.

Future research might explore how delays in receiving information about an event affect the control policies and effectiveness of the informed control strategy. Stochastic optimization could be used to capture the uncertainty in the timing, duration, and magnitude of events. The performance benefits observed in this chapter achieved with information about events are specific to holding control. The information might be more useful when other control strategies such as short-turning, dynamic deployment of reserve vehicles, and dynamic re-routing are available. For example, a greater performance improvement might be possible in the unforeseen link blockage case if vehicles were allowed to short-turn from the reverse peak direction to the peak direction downstream of the blockage. The applications presented in this research have a fixed dynamic demand that is independent of operator control actions and system performance. The framework could be applied to cases in which passengers can choose alternate services or modes in response to service alerts. In these cases demand models would play the critical role of forecasting the effect of events and service alerts on demand. The extraction, fusion, and interpretation of both historical and real-time data from multiple sources in the context of real-time transit control is a non-trivial problem with great potential for research.

Chapter 4

Schedule-Free Operations

High-frequency transit operations are subject to stochastic running times and demand that lead to differences between planned and delivered service. Researchers have proposed a number of real-time control strategies to maintain service quality and mitigate disruptions, most of which disregard schedules and aim for headway regularity. However, operations planning remains heavily focused on schedules, which are deterministic and constrain the availability of vehicles and crew. This dichotomy can sometimes put desirable control outcomes at odds with schedule constraints. The research presented in this chapter develops a schedule-free operations planning paradigm in which operations planning is driven by real-time optimization. Under the new paradigm, transit systems adapt to current and expected future conditions to maintain service quality while satisfying resource constraints.

The process of planning and delivering high-frequency public transportation service can be divided into three phases: service planning, operations planning, and service delivery. Service planning defines the service characteristics of importance to passengers, including network design and span and frequency of service. Operations planning determines how service will be delivered, generally as formalized in vehicle and crew schedules. Service delivery is the movement of vehicles and crew according to the operations plan, supplemented by control interventions to prevent and manage disruptions.

The planning and delivery process typically follows a *schedule-based paradigm*, under which the operations plan takes the form of a schedule and the principal aim of control in service delivery is schedule adherence. A drawback of this well-established paradigm is the dichotomy between a deterministic plan and a stochastic operating environment. Schedules specify planned stop times assuming particular running times, which may turn out to be shorter or longer once realized. This uncertainty is usually recognized and addressed through recovery time between trips, which provides a buffer that decreases the chance of lateness propagating between successive trips of a vehicle. Longer recovery times make schedules more robust to disruptions, but decrease fleet utilization and system capacity. Operators of high-demand systems often adopt aggressive schedules with less recovery time in order to increase capacity, making operations more susceptible to bunching and disruptions that are difficult to recover from.

Researchers have proposed a number of control strategies for high-frequency transit, and have shown that the most effective strategies do not adhere to the schedule, but instead aim for headway regularity or passenger cost minimization. These control strategies can generate policies that conflict with resource constraints. For example, holding a vehicle that is late with respect to the schedule might be desirable from a passenger cost perspective, but undesirable or impractical in a system with strict constraints on crew working hours. Existing control strategies neglect planned vehicle entries and exits, requiring operators to remain focused on the schedule, and perhaps discouraging them from embracing strategies that could help them improve service with the resources available.

This research proposes a *schedule-free paradigm* for high-frequency transit operations, in which operations planning is driven by optimization based on real-time information. In this paradigm, resources would be allocated before service delivery to a given service (or set of services), prespecifying only planned entry and exit times and locations for vehicles and crews. Unlike in the schedule-based paradigm, specific trips are not assigned to vehicles and crew beforehand. Instead, most operations planning decisions take place while service is being delivered, reflecting current and expected future conditions, and aiming for service quality while satisfying resource constraints.

The schedule-free paradigm enables a transit system to adapt recovery times, headways, and number of trips served to operating conditions as they exist. Flexibility is further increased when vehicles are shared among multiple lines, branches, or variations of a line. For example, short-turning can be used to increase frequency in the most heavily used portion of a line when overcrowding is detected (or expected). The sequence of trips served by each vehicle must allow the vehicle to meet exit constraints. A vehicle that does not have enough time to serve an additional round trip between terminals may be able to serve a short variation.

Apart from the supporting framework and models developed in this research, operating high-frequency transit without schedules is enabled by passengers' unawareness of schedules and recent advances in information technology. Passengers on high-frequency transit do not plan to take specific scheduled vehicle trips. Instead, they expect to wait a short time after their arrival at a stop (or station). Even when a vehicle schedule exists, passengers are typically not aware of it, since it is atypical for the operator to publish it. As long as their expectations of a short wait are met, passengers could enjoy improved service under the schedule-free paradigm without having to change their approach to using the service. Recent advances in information technology are another key enabling factor. Schedule-free operations planning relies on real-time sensing technologies to capture the current state of the system, powerful computing for plan optimization, and fast communications between vehicles and computers to transfer sensor data and update plans for near-term operations.

This research develops the framework and models that could be used to operate high-frequency transit without schedules, and evaluates the potential of the schedule-free paradigm for high-frequency transit. Section 4.1 motivates the research, Section 4.2 reviews the literature, Section 4.3 presents the framework, Section 4.4 discusses implications, Section 4.5 formulates the real-time trip planning problem and discusses its complexity and possible approaches, Section 4.6 presents a simple initial methodology, Section 4.7 applies the framework and methodology to a simulated transit service, and Section 4.8 draws conclusions.

4.1 Motivation

Under the schedule-based paradigm, schedules are generated with specific utility-generating service objectives in mind, considering waiting times, reliability, and crowding, among others. However, given a particular system state, the schedule may no longer be the best way of achieving those objectives going forward. Schedule adherence is an intermediate objective, since no utility is directly derived from following a schedule. The schedule-free paradigm pursues utility-generating objectives directly. Trip planning is driven by real-time optimization, with the goal of maximizing utility. Utility functions can reflect passenger cost components such as waiting time and in-vehicle time, as well as operator costs such as driver exit lateness. The function specification depends on the goals of the operator, as discussed in Section 4.3.

Transit operations are stochastic. Running times and demand change from day to day. Aside from the variability of running times due to traffic, signals, and driver behavior, and of dwell times due to different boarding speeds and fare media, there are low-probability events such as temporary lane blockages, wheelchair lift deployments, signal failures, and incidents that cause more significant delays. Under the schedule-based paradigm, operations are fully planned before service delivery, so even if the stochasticity of the system is measured and taken into account, the operations plan is rigid and deterministic. Aggressive schedules (with low recovery time) make more efficient use of vehicles when there are no delays, but even minor events can cause delays that are difficult to recover from without aggressive control actions such as short-turning and trip cancellations. Since late vehicles are not held longer than required for drivers to rest between trips, bunching ensues. As schedules are made more risk-averse by increasing recovery time, the tolerance for delays increases but fewer trips and lower line capacity are offered by the same fleet. Schedules are not adjusted during service delivery to reflect current and expected operating conditions. Conversely, the schedule-free paradigm defers planning the details of how vehicles will be used until service is being delivered, when the stochastic realizations leading to the present state have been observed and predictions of the short-term future are more certain because probability distributions of running times and demand can be conditioned on the observed current state and additional sources of information. Even when deterministic optimization methods drive schedule-free operations planning, short-term forecasts reflect the current state. Thus, schedule-free operations plans are generated with an information advantage over fixed schedules, and this advantage may lead to better performance. For example, higher frequency may be offered in the absence of delays, while holding time between trips can be increased to prevent bunching when facing delays. Some aspects of variability in a transit system can be predicted in the short-term. For example, weather may affect running times and demand, but it cannot be predicted far in advance. Where it might be impractical to adjust schedules based on weather forecasts, the operations plans generated under the new paradigm could consider the weather.

In addition to operating with reduced uncertainty, the schedule-free paradigm has the advantage of being able to explore many potential operations plans at any given moment. A schedule may, under particular conditions, be the optimal operations plan for a day. Even when this is the case, the same operations plan can be constructed through plan optimization under the schedule-free paradigm. The schedule-free paradigm has access to a potentially large set of feasible operations plans, one of which could be the traditional schedule (if one exists). Better plans can be generated and followed when the schedule is suboptimal. In principle, a transit service should perform at least as well under the schedule-free paradigm as under the schedule-based paradigm.

Some transit operators already temporarily abandon the schedule when facing large disruptions. For example, operators of London Buses may short-turn trips to absorb lateness. London Underground cancels trips during a disruption to ensure that crew exit constraints are met when recovery plans are generated. The operators of Santiago's metro short-turn trains to alleviate station congestion. These practices are typically employed based on experience, without supporting models, and with the intention of returning the system to schedule as quickly as possible. These examples suggest that operators, who already recognize the need to make exceptions to the schedule adherence objective in some circumstances, could find value in the schedule-free paradigm.

4.2 Literature Review

State-of-the-art operations control for high-frequency transit has advanced steadily over the past few decades, both methodologically and in terms of objectives. Control strategies such as holding, expressing, deadheading, and short-turning have been investigated, and increasingly rich decision support models have been proposed. The earliest models predate the availability of real-time vehicle location data (Osuna and Newell, 1972 and Barnett, 1974). More recent models utilize real-time data to capture the current state of the system and generate control policies accordingly (Eberlein et al., 2001, Daganzo and Pilachowski, 2011, and Bartholdi and Eisenstein, 2012). The most advanced models are based on rolling horizon optimization, which generate policies based on forecasts of system performance under potential control actions (Delgado et al., 2012, Sáez et al., 2012, and Chapter 2).

Throughout these advances there has been a move away from schedule adherence and toward headway adherence (Abkowitz and Lepofsky, 1990), headway regularity (Daganzo and Pilachowski, 2011, Cats et al., 2011, and Bartholdi and Eisenstein, 2012), and passenger cost minimization (Delgado et al., 2012, Sáez et al., 2012, and Chapter 2). The simplest control objective is schedule adherence, which aims to minimize deviations from the vehicle schedule. While this is a suitable aim for low frequency service in which passengers plan to take specific trips from the timetable and time their arrival at origin stops accordingly, researchers have long recognized that other strategies can improve performance in high-frequency transit when passenger arrivals are independent of the (often unpublished) timetable (Barnett, 1974).

In contrast to operations control, operations planning remains largely schedule-based. The schedule-based process of generating a timetable and vehicle and crew schedules is applied in the same manner to low-frequency and high-frequency transit, despite the differences in control objectives. Viewing operations planning and control for high-frequency transit together, current best practice is to produce schedules in the planning phase and subsequently ignore or abandon them in the service delivery phase to deal with disruptions. Operations planning for high-frequency transit has not yet evolved to become schedule-free.

Much of the work on real-time operations planning in transit has focused on disruptions management. A common and challenging problem is the recovery of a transit service after incidents cause delays and render the schedule infeasible. Among the works surveyed, the goal is invariably returning the system to the schedule as quickly as possible. Adenso-Díaz et al. (1999) and Şahin (1999) focus on minimizing changes to the original schedule. Walker et al. (2005) use integer programming to recover a train timetable and crew roster, minimizing deviations from the existing schedule and cost increase from adjusted crew shifts. Huisman and Wagelmans (2006) focus on real-time vehicle and crew scheduling given a timetable. Mazzarello and Ottaviani (2007) use heuristics to minimize delays of trains through a network of links, controlling speeds and considering re-routing. Törnquist and Persson (2007) address a similar problem with mixed integer linear programming, as do D'Ariano et al. (2007) using a discrete event model and a truncated branch and bound algorithm. Rodriguez (2007) uses constraint programming and a simulation model for real-time routing and scheduling of trains running through a junction. D'Ariano et al. (2008) test the concept of flexible timetables for railways, in which a timetable is generated in real-time to resolve conflicts, with the goal of minimizing delays with respect to the original (offline) timetable. Rezanova and Ryan (2010) focus on recovering the train driver schedule as soon as possible through the use of recovery time, re-routing, and trip cancellations. Corman et al. (2010) employ a tabu search algorithm for rerouting trains during disruptions with the goal of minimizing delays subject to resource constraints. Corman et al. (2012) minimize both train delays and missed connections (for passengers whose trips involve transfers). Krasemann (2012) combines a truncated branch and bound algorithm with guiding heuristics to obtain a quick response to incidents under scheduled service. Veelenturf et al. (2012) allow small delays in the timetable in exchange for greater flexibility in the real-time crew rescheduling problem, which results in fewer cancellations.

There has also been much work on service and operations planning before service delivery. The traditional process, which breaks the problem into a sequence of subproblems (frequency determination, timetable development, vehicle scheduling, and crew scheduling), is well established (Vuchic, 2005, Ceder, 2007, and Boyle, 2009). Desaulniers and Hickman (2007) survey operations research applications to service and operations planning. Recent developments have focused on increasing flexibility or integrating across the multi-step approach. Site and Filippi (1998) address the problem of service planning with short-turning and variable vehicle size. Leiva et al. (2010) optimize a combination of full and limited-stop services for an urban bus corridor with capacity constraints. Cortés et al. (2011) combine short-turning and deadheading for setting frequencies and vehicle capacities in a simple transit corridor. Valouxis and Housos (2002) combine bus and driver scheduling using heuristics and linear programming, focusing on scheduling bus service for the following day. Huisman (2007) develops a crew rescheduling model with the objective of minimizing cost subject to crew availability constraints, for situations in which changes to the timetable or vehicle schedule have made the original crew schedule infeasible, e.g. during infrastructure construction and repair works. Mesquita and Paias (2008) integrate vehicle and crew scheduling given a timetable combining a multicommodity network flow model with a set partitioning/covering model.

4.3 Framework

Transit service is planned in two stages: service planning and operations planning. Service plans define the transit network and service characteristics such as span of service and frequency, which influence both the kind of service passengers expect and the resources (for example, vehicles and drivers) required for operations. Operations plans define how an operator expects to deploy resources to deliver transit service to meet the service plan. In this stage operators analyze stochastic factors such as running time and demand variability and decide how many vehicles and drivers will be used for operations, trading off operational cost with risk of not being able to deliver the service characteristics defined in the service plan (for example, due to unexpected traffic delays in a bus corridor).

While service planning happens the same way under both schedule-based and schedule-free operations, there are significant differences in the way operations planning takes place. Under the schedule-based paradigm the operations plan is fully defined, and therefore fixed, before service delivery. Under the schedule-free paradigm operations planning begins before service delivery but mostly occurs in real-time reflecting the current system state and expected running times and demand. Figures 4-1 and 4-2 illustrate the two paradigms.

Service planning involves network design and service characterization under both paradigms.

Network Design The alignment of each line or route is defined. The objective is to connect different parts of the urban area to meet mobility and accessibility objectives. The locations of stations or stops are decided at this stage. Network design decisions have long-term implications that can influence demand and (over long periods) the urban landscape. For transit systems with dedicated ways and stations, the decisions result in infrastructure investments which are fixed. For bus lines without dedicated rights of way, the locations of stops can be changed more readily but it is still quite difficult. Politics, policy, and public involvement have significant influence at this stage.

Service Characterization The service characteristics of each line or route are defined. Span of service and frequency are set at this stage, responding to policy and expected demand. Service frequency typically varies by time of day. These decisions affect resource requirements and service standards. Higher frequencies increase the number of vehicles and crew required for operations, while decreasing waiting times and crowding for passengers. Service characteristics can influence demand because they affect accessibility and the relative convenience of a service with respect to alternative modes.

Under the schedule-based paradigm, operations planning involves timetable development and vehicle and crew scheduling.

Timetable Development Timetables specify vehicle departure times from stops or stations, reflecting the service frequencies set earlier as well as expected running times. At this stage trips are not yet assigned to vehicles. Although the schedule is usually published for low frequency service and passengers consult it to time their arrivals at stops, this is seldom the case for high-frequency transit. Therefore, the timetable is mostly an intermediate step in the planning process.

Vehicle Scheduling Vehicle scheduling assigns sequences of trips from the timetable to specific vehicles, resulting in the sets of trips to be served by each vehicle. The required fleet size is determined based on the timetable, rules

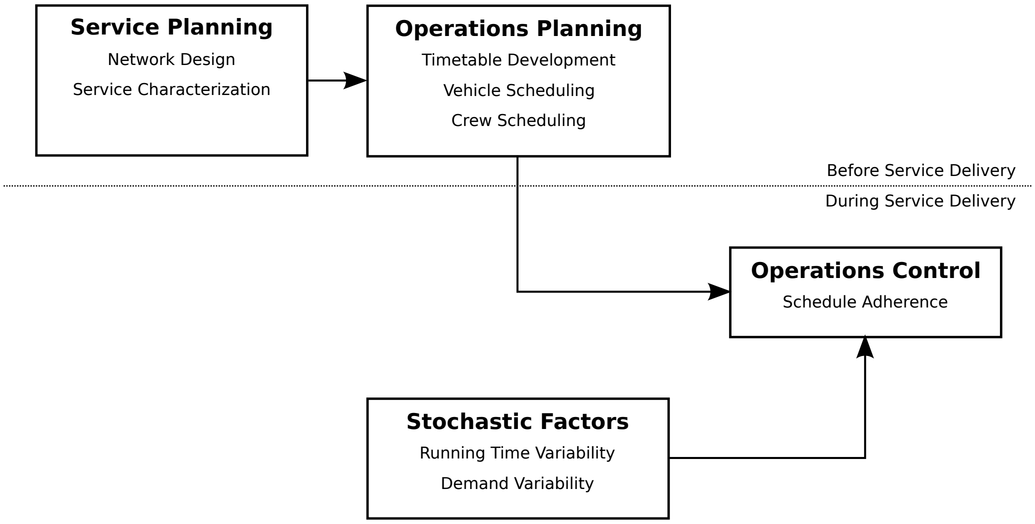


Figure 4-1: Schedule-Based Paradigm

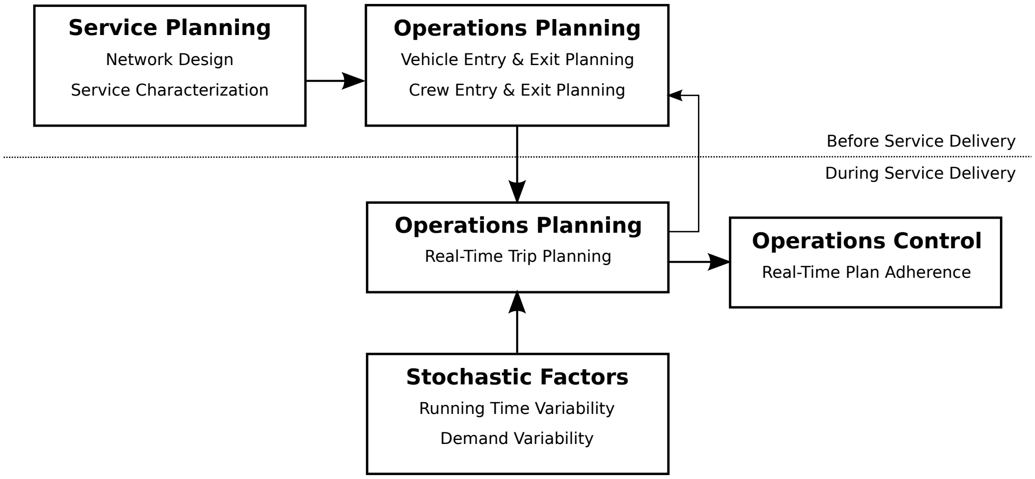


Figure 4-2: Schedule-Free Paradigm

governing minimum layover times at terminals, and running times. Recovery time is added to expected end-to-end running times so that longer than expected running times do not cause lateness to propagate across trips. Recovery time increases the robustness of schedules but also the required fleet size. The number of active vehicles typically varies by time of day. Minor adjustments to the timetable are often made in order to decrease the number of required vehicles. The primary objective is minimizing fleet size in order to reduce operating cost.

Crew Scheduling Crew scheduling assigns sets of vehicle trips to drivers in accordance with work rules governing shift durations, breaks, and pay provisions. Typically, multiple drivers are required to cover the trips served by each vehicle. Minor adjustments to the timetable and the vehicle schedule may be made in order to decrease the number of required drivers while satisfying all work rules. The primary objective is minimizing operating cost. Crew scheduling is not required for autonomous vehicle fleets.

Under the schedule-based paradigm, all planning decisions take place before service delivery. During service delivery, operations control focuses on schedule adherence, which is meant to result in service that meets service planning objectives. Vehicles are held at terminals and other control locations to prevent early departures, and depart as soon as possible after late arrivals. Stochastic factors affect operations, sometimes causing delays and overcrowding, but there is no provision to adjust the plan to reflect current and expected operating conditions.

The schedule-free paradigm defers some operations planning decisions until service delivery. Instead of planning the deployment of vehicles and crew at the trip and stop level, only their entry and exit times are planned before service delivery. Trip and stop level activities are planned during service delivery, which allows planned service to adapt to current and expected conditions, utilizing observations of stochastic running time and demand realizations that are not available before service delivery.

Vehicle and Crew Entry and Exit Planning Entry and exit plans specify when and where vehicles and crew enter and exit service, but not the specific set of trips each vehicle and driver serves. Entry times define the earliest allowed planned dispatch, while exit times define the latest allowed planned end of a trip. The number of active vehicles and crew, which can vary by time of day, should reflect the frequencies of the service plan as well as running times and demand. Vehicle and crew availability must be decided before service delivery because drivers need to know their check-in and check-out times and agencies need to allocate resources to routes and budget operations. Operators may assign vehicles and drivers to a single line, or they may allow them to be used across multiple lines (for example, a set of lines sharing a terminal).

Real-Time Trip Planning The specific trips each vehicle and driver serve are planned in real-time during service delivery, and adjusted based on operating conditions, aiming to minimize a combination of passenger and operator costs while satisfying the constraints defined in the entry and exit plan. Real-time planning must consider the time remaining until each vehicle's and driver's latest allowed planned exit and the feasibility and cost of completing each vehicle's planned sequence of trips. The output of the optimization-driven process specifies target departure times for all planned stop visits. These plans, which can be updated every few minutes, are communicated to vehicles in real-time and treated like a schedule for control purposes. The availability of a vehicle according to the entry and exit plan does not require the real-time plan to use it. For example, some vehicles may be reserved to respond to disruptions, the decision for the vehicle to enter being part of the real-time planning process.

Since the operations plan can be adjusted in real-time, it reflects operating conditions influenced by stochastic factors such as running time and demand. Consequently, the operations plan remains realistic with respect to the current and expected system states, and reflects objectives of minimizing passenger and operator costs and satisfying constraints. As a result, there is no longer a contradiction between control policies that improve adherence to the operations plan and those that improve service for passengers.

Real-time trip planning involves complex decisions to coordinate vehicle and driver movement in a way that leads to good service for passengers but avoids excessive operator costs or constraint violations. For example, a particular vehicle may not have enough time left before its planned exit to serve an additional two round trips between terminals due to previous delays. One option is to serve both trips and incur an exit lateness cost, if the work rules allow it. Another option is to serve a single trip, either exiting earlier than planned or holding at the terminal longer than normal before starting the last trip. Yet another option is to serve one round trip between terminals and a second short trip. Either the first or second round trip could be short-turned. Each of these options may have different implications for passenger and operator costs, as well as on the optimal trip sequences of other vehicles. For example, if a pair of vehicles is bunched and both must be short-turned in order to satisfy exit constraints, it may be advantageous to short-turn them on different trips, so that a long headway does not result from two consecutive vehicles being short-turned. Departure times from terminals and stops must be optimized in addition to trip sequences, and these are interconnected. Trip plans with little or no slack before the latest allowed exit time do not allow for additional holding time to be added later, thus limiting the ability to respond to future disruptions. The optimization problem becomes large and complex when the trip sequences of an entire line's fleet are being considered.

Strategies employed to meet the operations objective may include holding, short-turning, dead-heading, expressing, and injecting reserve vehicles. Holding delays vehicles while short-turning, dead-heading, and expressing advance vehicles. Expressing and short-turning trips may force some passengers to alight and board a different vehicle to complete their trips. Some passengers may choose to complete their journey on foot. These strategies impose additional costs on passengers in terms of waiting time, trip time, reliability, and convenience, leading to degraded overall service quality, but they can be employed during incidents and severe disruptions to restore service when the alternatives are worse. Dead-heading and short-turning decided at the terminal are applied only to empty vehicles, sparing passengers the frustration of having to change vehicles.

The schedule-free paradigm allows real-time interlining between high-frequency lines sharing a terminal. Vehicles and crew can be assigned to serve more than one transit line in the vehicle and crew entry and exit plan. Resources are then assigned to lines on a trip by trip basis through the real-time trip planning optimization process. When a vehicle arrives at the shared terminal, it can be dispatched to the line that best serves the cost minimization objective. When cycle times differ, interlining can be used like short-turning to increase vehicle utilization while complying with exit constraints. For example, a delayed vehicle may arrive at a terminal without enough time to serve an additional cycle on its usual route, but with enough time to serve a cycle on a shorter route; this option is better than taking the vehicle out of service much earlier than its latest allowed exit time. In schedule-based operations, the same vehicle would continue serving its next (delayed) scheduled trip, possibly leading to a late exit.

Interlining under the schedule-free paradigm serves the purpose of improving service for passengers and increasing vehicle utilization, rather than reducing fleet size, which is the usual aim of interlining low-frequency transit services. Interlining to reduce fleet size is less attractive for high-frequency transit (under both paradigms) because high-frequency lines are more prone to headway imbalances, and because potential vehicle savings are equivalent to shorter extra recovery times (without interlining), making it more likely for delays to propagate between lines. Table 4.1 shows an example of the effects of interlining two lines, with cycle times of 70 and 50 minutes, for low-frequency and high-frequency service. Interlining saves 1 vehicle in both cases, but the extra cycle time without interlining, which is the waste of resources that interlining can prevent, is 10 minutes for the low-frequency service and 2 minutes for the high-frequency service. Assuming that the cycle times in this example are normally distributed and that target cycle times correspond to 90$^{\text{th}}$ percentile cycle times, the percentiles corresponding to the cycle times without interlining are 99 in in the low-frequency case and 94 in the high-frequency case. The operator is more likely to find the former excessive.

Entry and exit plans can reflect different cost structures and objectives, from tight exit constraints under strict work rules to fixed unit operational cost without hard exit constraints for driverless fleets. Exit constraints can be a combination of strict and flexible. Strict entry times might reflect the

Table 4.1: Interlining Low-Frequency and High-Frequency Transit

	low frequency		high frequency	
headway	20 min		4 min	
typical cycle time	60 min	40 min	60 min	40 min
target recovery time	10 min	10 min	10 min	10 min
target cycle time	70 min	50 min	70 min	50 min
fleet size w/o interlining	4 veh	3 veh	18 veh	13 veh
combined fleet size w/o interlining	7 veh		31 veh	
combined cycle time	120 min		120 min	
combined fleet size w/ interlining	6 veh		30 veh	
fleet savings	1 veh		1 veh	
cycle time w/o interlining	80 min	60 min	72 min	52 min
recovery time w/o interlining	20 min	20 min	12 min	12 min

check-in times of drivers, which may not be altered in real-time. Flexible exit times might reflect the desire for a driver to exit by a certain time with a possibility for overtime payment for the time served after the planned exit time, in cases where some exit lateness can be traded off with better service quality for passengers. Both types of constraints may be used simultaneously; for example, driver lateness of up to 30 minutes might be allowed at a cost, and a strict exit constraint can reflect the infeasibility of trip plans leading to exit lateness greater than 30 minutes. Costs associated with flexible resource constraints can increase nonlinearly with lateness; for example, lateness of 20 minutes might be considered more than 4 times as costly as lateness of 5 minutes.

A transit system's performance under the schedule-free paradigm feeds back into entry and exit planning. For instance, if overcrowding or excessive waiting is observed, an operator can allocate additional vehicles when required. Simulation can be used to predict performance with a given entry and exit plan when no observations of real service exist or when systematic changes in the operating environment are expected, for example due to change of season, change of the transit network, or events. The traditional schedule-based approach can be used to make a first entry and exit plan if one does not exist, keeping the times at which vehicles and crews enter and exit service without defining trip-level detail. Entry and exit plans limit what can be achieved in real-time, so it can be beneficial to optimize them. Simulations of different entry and exit plans can drive their optimization.

Overly tight constraints on entries and exits can significantly limit the flexibility and performance of schedule-free operations. For example, in a schedule-based service with long cycle times, there may only be enough time for a vehicle to serve a single round trip, without much time for holding. In this case the constraints only allow the vehicle to be dispatched when it enters the system and to run as fast as possible to its exit. The same outcome is expected under the schedule-free paradigm. Adding buffer time before scheduled exit times gives the real-time planner under the schedule-free paradigm the flexibility it needs to optimize dispatch times and trip sequences. While in the schedule-based paradigm specific amounts of buffer time are added between (scheduled) trips, in the schedule-free paradigm buffer time is not planned for specific times, but rather generally available for optimized usage.

Schedule-free operations may not always result in more trips offered with a given fleet than schedule-based operations. Given enough buffer time, the schedule-free paradigm can increase headways when facing longer running times in order to prevent bunching. The required holding might happen during delays, or later if there is a capacity problem and it is better to postpone holding. This can increase vehicle utilization, i.e. the total passenger benefit obtained from vehicles, with similar usage, i.e. number of trips offered, or total duration (or distance) of runs. However, it may be difficult to improve efficiency (with respect to schedule-based operations) when scheduled cycle times are higher than realized cycle times, because the total excess recovery time in the schedule may not be sufficient to offer additional trips.

For example, consider a schedule having 5 more minutes of recovery time after every (one-way) trip than what is required to regulate headways, such that actual cycles can be completed in 70 minutes, while the scheduled cycle time is 80 minutes. If the entry and exit constraints give enough time for only 3 (80 minute) round trips (i.e. 6 one-way trips), then the total extra recovery time is 30 minutes. This is insufficient to serve another full cycle, and possibly a short cycle. In this case the schedule-free plan should be to run 3 cycles, but the timing of trips must be decided. If vehicles are not held more than necessary to regulate headways, there will at first be higher frequency, but then vehicles will be taken out of service 30 minutes early, leading to lower frequency and possibly disrupted service. It is possible, and perhaps (depending on demand) likely, that the best plan in this case is to hold vehicles to run cycles of 80 minutes, as in the schedule, in which case the 30 minutes of extra time are wasted under both paradigms.

Driverless systems might not limit how long vehicles can operate for, or their entry and exit times, but only the maximum number of vehicles dedicated to a service or set of services. In this case the operator assumes a fixed unit cost. Real-time planning might consider the trade-off between this cost and passenger cost, dispatching a vehicle if passenger benefits exceed operating costs. Policies and service standards can also drive trip planning. For example, a policy headway of 15 minutes might trigger a dispatch late at night even if marginal operating costs exceed expected marginal passenger benefits.

There are cases in which transit services are partly high-frequency and partly low-frequency. Many high-frequency services operate at low-frequency early in the morning and late at night, periods during which passengers may follow a schedule. The schedule-based and schedule-free paradigms may be combined, such that a schedule is followed during low frequency operations and the service switches to schedule-free during high-frequency operations. The transition from schedule-based to schedule-free is simple because schedule-free real-time planning can take place regardless of the current state. The transition from schedule-free to schedule-based is more complex because vehicles must be available at particular times to serve scheduled trips. A smooth transition can be facilitated by an entry and exit plan that specifies the times and locations at which vehicles begin operating schedule-based, in the same way exit times and locations are specified. This encourages real-time plans that make vehicles available to the schedule-based paradigm in time for each vehicle's first scheduled trip.

Some branched transit systems operate with high frequency on the trunk portion but with low frequency on the branches. Passengers originating in the trunk portion and destined for a branch stop usually wait for the first vehicle serving their branch. The schedule-based operation in the branches constrains what can be achieved with real-time operations planning in the trunk, making it difficult to avoid schedules entirely. However, if a subset of vehicles is dedicated to high-frequency operation on the trunk portion or perhaps on branches operating in high-frequency, then schedule-free real-time trip planning could improve service with optimization that reflects the trajectories of vehicles following a schedule.

Real-time trip planning can be made robust to temporary interruptions in communication between the planning computer server and vehicles. Trip plans include instructions for operating during a sufficiently long period after the plan is generated, so drivers can follow the plan most recently received. If communication with a vehicle is interrupted, the planning computer can generate plans taking into account that the disconnected vehicle will be following an older trip plan.

4.4 Implications

Adopting the schedule-free paradigm has implications on many aspects of transit service provision. In current practice, schedules play a role that extends beyond operations planning, to performance measurement, incident response, and passenger information provision.

4.4.1 Reliance on Information Technologies

Transit operators have increased the use of information technologies to plan, control, and measure the performance of service over the past decade. Automated data collection systems allow operators to measure demand and running times to make better schedules. While these technologies are already important, they are not essential, because schedule-based service can be delivered without them.

The new paradigm requires reliable and robust communications and information technologies, much more so than for schedule-based operations. While scheduled service can be delivered with only printed schedules for drivers to follow, schedule-free operations require frequent communication between vehicles and the computer optimizing plans in real-time. Vehicles must be equipped with automated vehicle location systems that transmit their location to the server handling real-time planning and control. It is useful, but not necessary, for them to have automated fare collection and automatic passenger counting in order to collect data that can be used to model running times and demand from historical data. It is also useful, but not necessary, for fare collection and passenger counting data to be transmitted to the agency in real-time so that the current state can be estimated more accurately than with historical data alone. In schedule-free operations, plans governing when a vehicle should hold, dead-head, or short-turn, and what its destination sign should display, can be updated every few minutes to reflect new information about current and expected conditions, so it is imperative that vehicles can receive updated plans on the go and display them to the driver.

Successful real-time planning may require sophisticated forecasting capabilities that utilize automatically collected vehicle location, fare collection, and passenger counting data to model running times and demand dynamically. Significant computing resources may have to be dedicated to optimization in order to operate many high-frequency services in real-time.

4.4.2 Methods and Practices

The schedule-free paradigm changes how an operator plans operations, responds to incidents and disruptions, measures performance, and provides information to passengers. Fundamentally, operations planning changes from something that happens entirely before service delivery to something that happens mostly during service delivery. Only entries and exits are planned a-priori, while trip and stop level details are planned in real-time and can change from day to day. However, there are other implications on methods and practices.

The schedule-free paradigm changes how transit service providers respond to incidents. Incidents such as signal failures or medical emergencies delay vehicles and result in disrupted service: imbalanced headways, crowding, and long waiting and trip times, which can last far longer than the incident itself. In scheduled operations, it is often necessary to abandon the schedule at least temporarily and use holding, short-turning, dead-heading, and injection of reserve vehicles to restore normal service. The objectives motivating these aggressive control actions are principally to reduce overcrowding and to minimize driver lateness at the end of duties. When faced with severe disruptions, controllers may cancel trips and reassign scheduled trips to different vehicles. This type of real-time adjustment of an operations plan is often conducted manually based on experience and heuristics, although increasingly with the aid of decision support systems. This creates an artificial divide between normal operations and disrupted operations, with the vehicle schedule serving as a measuring stick for disruption magnitude. Responding to incidents can occupy controllers, leaving no one to supervise and control the rest of the system. This lack of attention due to lack of human resources and support technology can negatively affect services not directly affected by an incident.

In schedule-free operations, there is no hard line separating normal operations from disrupted operations. Current and expected conditions together lead to different operations plans. When an incident occurs, information about it can (optionally) be supplied to the optimization model to improve the accuracy of its forecasts. The same methods used to optimize plans during normal operations generate policies reflecting the incident. Incidents requiring movement of crew with shuttles or emergency bus services still require manual coordination. Computer optimized real-time operations planning replaces the manual process for routine disruptions, leaving staff more time to focus on resolving the incident causing the disruption and on major disruptions requiring more complex responses. Since human resources are not involved in routine control tasks, services not directly affected by an incident continue to receive the same amount of planning and control attention.

The schedule-free paradigm makes it more important for transit agencies to use passenger-centric performance measures for high-frequency transit. Without a schedule, there is no notion of punctuality. Excess waiting times can be determined based on headways from the service plan. Estimated waiting times, trip times, and loads can be compared with service standards.

The schedule-free paradigm offers a unified approach in which real-time operations planning, performance measurement, and passenger information provision happen simultaneously based on the same information. The same model used to generate plans in real-time can be used to measure performance and provide information to passengers. Detailed vehicle location and demand data, including fare collection and passenger counting data, if available, feed into the model. The model estimates current and future system states, including arrival times, loads, waiting times, and in-vehicle times. The different elements can be combined into a generalized passenger cost performance measure. Waiting time statistics provided by the model would be more accurate than their simpler headway-based counterparts. Performance measures based on current system state can be stored for ex-post analysis, while estimates of future performance can be shared with supervisors, controllers, and managers in real-time. Information can also be shared with passengers through a variety of channels including displays, announcements, web pages, and mobile phone applications. Estimates of arrival times for each vehicle can be used to provide users with the next arrival time at each stop. In addition, passengers can be shown estimated loads and estimated arrival time at destination (e.g. if they state their destination in a web page or mobile phone application). This information is a fully automated product of the model.

4.4.3 Human Factors

Since passengers of high-frequency transit do not follow a schedule, they can continue using the service without changing their approach or even being aware of the new paradigm. The change to real-time planning directly based on maximizing passenger benefit can result in lower waiting times and improved reliability. Depending on current practice, stop-skipping strategies such as short-turning, deadheading, and expressing might be employed more frequently. Although in their benign forms they do not force passengers to alight before reaching their destination, passengers may experience a greater number of vehicles passing their origin stop but not serving their destination, in particular with deadheading and expressing.

The experience of drivers may change more significantly. If a high exit lateness cost is assumed for operations planning optimization, drivers might be able to expect to finish their duties by their target end time more reliably during severe delays, as long as real-time plans reflect delays. Differences between planned and real running times, which could be due to modeling errors or unexpected delays, could cause late exits regardless. For example, drivers delayed in their last cycle might exit late if the plan when the cycle begins does not reflect (future) delays and has insufficient buffer time between the planned and latest allowed exit times. In fact, exit lateness could increase (with respect to schedule-based operations) when a service experiences unexpected delays for a prolonged period, because holding, which may be used to regulate headways after delays are first seen, may lead to real-time plans having vehicles exit closer to the latest allowed exits. Additional (unexpected) delays can later render these plans infeasible, and for some vehicles (particularly those in their last cycle), there may not be feasible alternatives. This is less likely under schedule-based operations because delayed vehicles are not held (beyond required layover times for driver rest) when delayed. Drivers may experience less certainty about the sequence of trips they will carry out, because the plan may change throughout the day in response to changing demand and running times.

When current operations are based on an unrealistic schedule, drivers can often be late with respect to the schedule. This can cause frustration and stress, and eventually result in disregard and dismissal of the operations plan. Under the schedule-free paradigm, the plan adjusts to current and expected conditions, remaining realistic, and making it easier for drivers to follow the operations plan. Schedules make it possible for perverse and hidden incentives to develop. For example, labor union agreements often constrain shift lengths which may encourage the operator to hire more drivers rather than planning longer shifts, but individual drivers may want overtime compensation. This can motivate operators to generate unrealistic schedules that meet the constraints of the labor agreement on paper while giving drivers overtime. Overtime can be expected in the schedule-free paradigm only when drivers incur unexpected delays towards the end of their shift, or if the allowed duration of shifts explicitly set in the entry and exit plan is long enough to allow overtime. The schedule-free paradigm can also lead to perverse incentives, because knowing that plans will adapt, drivers and managers may put less effort into following the operations plan.

The performance of a transit line depends on driver compliance under both paradigms, but service may be more robust to a limited amount of non-compliance under the schedule-free paradigm. A system that sends control instructions automatically can log instructions and maintain a record of compliance for each driver, which facilitates enforcement and encourages higher compliance rates. Although this bookkeeping is possible under both paradigms, it is more likely to be implemented in a schedule-free system because the technology to send updated dispatch times to drivers and actively monitor vehicle locations is necessary.

The experience of controllers responsible for managing disruptions may also change. Since the schedule-free paradigm adapts to current and expected conditions, routine minor disruptions might be automatically handled by computer-based real-time planning. This can be accomplished, to some extent, under schedule-based operations using control models, but controllers may have to adjust the strategies to satisfy resource constraints, so their attention may still be required. The schedule-free planning algorithm could be better able to maintain high service quality while satisfying constraints, potentially decreasing required human involvement. The planning algorithm may lead to more frequent stop-skipping instructions, potentially making it difficult for controllers to keep track of the plans being considered. Implementing schedule-free operations may be difficult without automatic ways of communicating plan changes to drivers (as assumed in this research), because controllers might be overwhelmed with instructions to communicate. However, implemented with automatic handling of routine disruptions and plan updates, the schedule-free paradigm could allow controllers to focus on situations requiring creative or complex response strategies, such as disabled vehicles, signals malfunctioning, or incidents that require dispatching reserve vehicles and drivers, re-routing, introducing temporary replacement service, etc. A schedule-free planning model could interactively support these tasks by predicting performance under the various response options being considered, and by automatically generating the remainder of the operations plan in light of manual decisions. For example, a (human) controller might respond to a major disruption by short-turning some vehicles, and the schedule-free plan optimization algorithm might assist by optimizing the remainder of the plan. Operating on its own, the planning algorithm could alert controllers when service quality decreases (or is predicted to decrease) below established thresholds.

4.4.4 Operations with Uncertain Resources

The schedule-free paradigm could be attractive to local governments seeking to organize informal public transport systems, in which fleets or even single vehicles are independently owned by drivers. Minimum technical requirements for vehicle hardware could be fulfilled with Internet-enabled mobile phones, and cloud computing services could be used for server-side planning and optimization. These options are available in most parts of the world, making implementation technically possible in developing countries. While it may be difficult to know far in advance how many vehicles will be available for operation of a transit corridor, the new paradigm would allow drivers to check in with their phones and begin offering service with little lead time. The system would generate real-time plans that coordinate the movement of independent vehicles to serve the corridor efficiently. Drivers could also specify their target time for finishing work, so that the system could capture planned exits as plans are generated. Drivers could be compensated based on time worked, number of trips, and compliance with the operations plan. Economic incentives could be used to allocate drivers (voluntarily) to transit corridors needing more vehicles. Challenges include obtaining commitment from drivers to serve trips as planned and ensuring that vehicles are actually serving passengers, rather than driving in the corridor without stopping to allow passengers to board.

4.4.5 Operations with Autonomous Fleets

Autonomous vehicles are the future of high-frequency transit, because they can operate at higher frequency, lower costs, and with fewer constraints. Technologies for autonomous driving have been advancing rapidly. In 2013 there were 48 automated metro lines (without staff on board) in operation in 32 cities (UITP, 2013). The schedule-based paradigm places unnecessary constraints on a transit line with a driverless fleet, reducing its potential by restricting flexibility. The schedule-free paradigm can take advantage of the added flexibility, and can also accommodate a mixed fleet having some vehicles driven by human drivers and others operating autonomously.

Schedule-free operations with autonomous vehicles are less constrained than those with drivers, because there are no specific times by which vehicles must exit. Although humans may be supervising operations from a control center, there are usually fewer supervisors than drivers, often having less strict work rules, making it less costly to provide for the possibility of operations ending later some days. Without human drivers, only fleet size must be decided before service delivery. Although there are no driver constraints, it may be undesirable to have all vehicles working all the time, because there are operating expenses such as energy and maintenance costs that are proportional to time operated and distance run. Real-time planning can be based on the headways defined in the service plan, so that vehicles are brought into the system as frequency increases and are taken out of service as frequency decreases, including taking the whole fleet out of service after the planned service end time. Service standards can also contribute to these real-time decisions to pull vehicles in and out of service. For example, estimated future loads can be compared with the maximum loads in the service standard, and additional vehicles can be brought into service to manage overcrowding.

4.5 General Methodology

Schedule-free transit is driven by real-time operations plan optimization. Plans define the future trajectory of each vehicle from its current or future entry location to its exit, specifying the sequence of stop visits with target arrival and departure times.

Figure 4-3 illustrates the schedule-free operations architecture. The controller uses the dynamically modeled running times and demand to maintain an estimate of the current state. Every time a vehicle visits a stop, the estimated number of passengers inside and number of passengers left behind at the stop (by origin-destination pair) are updated. The operations plan is consulted to determine planned departure times; vehicles hold if they are ahead of the planned trajectory and holding is allowed at the current location. Vehicles may skip stops through strategies such as short-turning and deadheading as dictated by the plan. The plan is updated at regular intervals, e.g. every 5 minutes.

The first step in the process to update the operations plan is modeling the current state of the system, which sets boundary conditions for the subsequent plan optimization step. The current state includes locations of vehicles in the system, each vehicle's variation, the number of passengers in vehicles (by destination), the number of passengers waiting at each stop (by destination), the previous vehicle departure time from each stop, the current vehicle or location of drivers in the system, and the (planned) entry times and locations of vehicles and drivers not yet in the system. These are inputs to the plan optimizer, along with minimum and maximum holding times by stop, dynamic running time and demand functions, unit boarding and alighting times per passenger, weights for passenger waiting time, in-vehicle time, and driver exit lateness, and scheduled exit times and locations for vehicles and drivers.

Real-time operations plans are generated based on these inputs by opti-

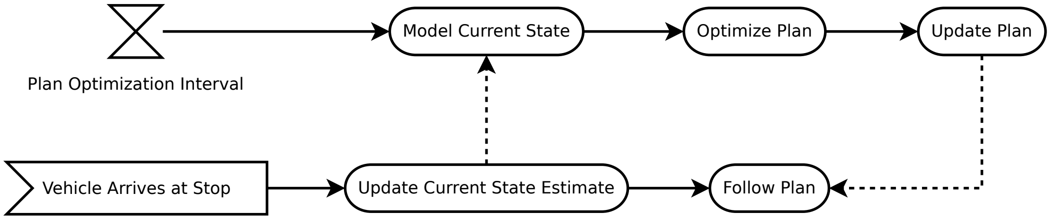


Figure 4-3: Schedule-Free Operations Architecture

mizing *trip sequences* (the spatial dimension) and *departure times* (the temporal dimension), which together define vehicle *trajectories*. The collection of planned trajectories (for all vehicles) defines a *plan*, with corresponding headways, loads, passenger waiting and in-vehicle times, and vehicle exit times. Departure times are driven by holding, running times, dwell times, etc. The objective is to minimize a combination of passenger and operator costs while meeting resource constraints. For simplicity, vehicles and drivers are considered a single entity; drivers are not explicitly modeled, but their entry and exit constraints are assigned to the vehicle they operate. This assumption holds for transit lines in which drivers take the vehicles in and out of service, and there are sufficient vehicles for each driver. Due to this simplification, the methodology developed in this section cannot directly capture en-route driver reliefs or constraints relating multiple pieces of work by the same driver, e.g. minimum meal break durations. Constraints modeling real-time assignment of drivers to vehicles (or vice versa) would be required to accomplish this. The trip plan optimization problem can be formulated as

$$\begin{aligned} \underset{x \in X}{\text{minimize}} \quad & C(x;p) && (4.1)\\ \text{subject to} \quad & u_v \leq u''_v \quad \forall v \in V && (4.2)\\ & z_v = z'_v \quad \forall v \in V && (4.3)\\ & \text{vehicle movement constraints} && (4.4)\\ & \text{passenger activity constraints} && (4.5)\\ & h_e^{\min} \leq h_e \leq h_e^{\max} \quad \forall e \in E && (4.6) \end{aligned}$$

where x is a candidate plan, X is the set of feasible plans, p is a set of exogenous parameters and initial conditions, u_v and u''_v are the planned and latest allowed exit times of vehicle v, z_v and z'_v are the planned and required exit locations of vehicle v, h_e is the holding time corresponding to a planned stop visit e, $h_e^{\min}$ and $h_e^{\max}$ are lower and upper bounds on holding times at the same planned stop visit, V is the set of vehicles, E is the set of planned vehicle stop visits, and $C(x;p)$ is a general non-convex cost function covering the modeling horizon, subject to general constraints that, in addition to those explicitly listed, define initial conditions, and vehicle and passenger movement. The variables defining a plan x are both continuous (departure times) and discrete (trip and stop sequences). Constraints (4.2) and (4.3) ensure that plans deliver vehicles to their exit locations by the required times, while constraint (4.6) limits holding times at terminals, turning points, and stops. Terminals and en-route turning points are modeled as stops without demand.

The cost measure C combines mean passenger cost C_P, exit lateness cost C_L, and plan complexity cost C_C:

$$C = C_P + \theta_L C_L + \theta_C C_C \tag{4.7}$$

where θ_L and θ_C are the relative weights of exit lateness and complexity, respectively.

Passenger cost captures waiting time at stops and in-vehicle time, over a horizon extending from the current time t_0 to t_f, for a given horizon length $t_f - t_0$. It includes the in-horizon portion of waiting time W_f for passengers who are still waiting at the end of the horizon. A discount factor can be applied to weight costs incurred sooner more heavily than costs incurred later. This reflects growing uncertainty of predicted future states over time. While deterministic predictions are fairly accurate for states in the near future, stochasticity and model inaccuracies may lead to prediction errors that accumulate over time, potentially leading to significant differences between actual and predicted costs. The discount factor is of the form $\mathrm{e}^{\beta(t-t_0)} \in [0,1] \, ; \beta \leq 0$. Mean passenger cost is given by

$$C_P = \frac{\sum_{i=1}^{n} \mathrm{e}^{\beta(t_i - t_0)} \left(V_i + \theta_W W_i\right) + \mathrm{e}^{\beta\left(t_f - t_0\right)} \theta_W W_f}{P} \tag{4.8}$$

where n is the number of planned stop visits, t_i, V_i and W_i denote the departure time, in-vehicle cost, and waiting cost of the i^{th} planned stop visit, respectively, t_f and W_f denote the time and waiting cost at the end of the horizon, respectively, θ_W is the relative disutility of passenger waiting time, and P is the total number of boardings.

Exit lateness cost can be a general function. We use the following piecewise-polynomial specification:

$$C_L = \sum_{v \in V} \max\Big(0, \left(u_v - u_v'\right)^{\alpha}\Big) \tag{4.9}$$

where u_v and u_v' are the planned and target exit times of vehicle v, and α is a constant parameter. *Planned exits* are those resulting from a candidate operations plan x. *Target exit times* come from vehicle and crew entry and exit plans, determined before service delivery. Real-time operations plans can use vehicles and drivers up to their target exit times without lateness cost. Some lateness may be allowed at a cost, but exits later than u'' may not be planned. By construction, target exit times u' must not be later than latest allowed exit times u''. Values $\alpha > 1$ can be adopted as a disincentive for very late exits. With discounting it becomes

$$C_L = \sum_{v \in V} \max\Big(0, \mathrm{e}^{\beta(u_v - t_0)} \left(u_v - u_v'\right)^{\alpha}\Big) \tag{4.10}$$

Plan complexity cost C_C can be added as a disincentive for plans requiring a lot of stop-skipping (e.g. short-turning) or holding at many stops for only marginal performance improvements. For example, complexity may be a function of the number of planned short-turns.

The planning problem can be decomposed, without loss of generality, into a *trip sequences* problem and a *departure times* subproblem. This decomposition is natural because trip sequences are discrete while departure times are continuous. Mathematically, the trip sequence problem is

$$\underset{s \in S}{\text{minimize}} \quad C(s; d_s^*, p) \tag{4.11}$$

where s is a candidate combination of trip sequences for all vehicles, S is the set of all feasible trip sequence combinations, and d_s^* are optimal departure times for each given trip sequence combination s, provided by the departure time subproblem:

$$\underset{d \in D}{\text{minimize}} \quad C(d; s, p) \tag{4.12}$$

where d is a set of departure times for all vehicles, D denotes the feasible space of departure times, and s is a candidate combination of trip sequences for all vehicles, given by the master problem. Constraints (4.2) through (4.6) apply, as before, in both the master problem and the subproblem.

4.5.1 Problem Complexity

The schedule-free real-time planning problem is large, complex, and difficult to solve. Changes to a planned stop visit, in terms of either location or timing, affect following planned stop visits for the same vehicle, including the number of passengers waiting at the stop, dwell time, feasible departure times, and possible next stops. Departure times, in turn, affect running times, and therefore future stop visits, and exit times, which determine whether a trip sequence is feasible. Trip sequences of one vehicle affect those of nearby upstream vehicles through the number of passengers waiting, the order in which vehicles visit stops, etc. Trip sequences are inherently discrete, making it difficult to model mathematically the relationship between alternative trip sequences for a particular vehicle in terms of expected cost differences. In addition, the costs of candidate trip sequences are highly dependent on departure times, because they determine headways, waiting times, and exit lateness. This makes it difficult to estimate the benefits of different trip sequences without first optimizing departure times.

Ideally, all trip sequences and departure times would be optimized together. Unfortunately, this problem grows combinatorially, making it impractical to solve in real-time. Its complexity is

$$\mathcal{O}\left(\sum_{i=1}^{\prod_{v \in V} |S_v|} k_i\right) \tag{4.13}$$

where v is a vehicle, V is the set of all vehicles, S_v is the set of candidate trip sequences for vehicle v, $|S_v|$ is the set's cardinality, i.e. the number of candidate trip sequences, and k_i is the complexity of optimizing departure times for a set of trip sequences i. Assuming a constant complexity k for departure time optimization, the complexity of the plan optimization problem simplifies to

$$\mathcal{O}\left(\sum_{i=1}^{\prod_{v\in V}|S_v|} k\right) = \mathcal{O}\left(k\prod_{v\in V}|S_v|\right) \tag{4.14}$$

Assuming a constant number of candidate trip sequences n for each vehicle, complexity further simplifies to

$$\mathcal{O}\left(k\prod_{v\in V} n\right) = \mathcal{O}\left(kn^{|V|}\right) \tag{4.15}$$

where $|V|$ is the number of vehicles. These simplifications are only for illustrative purposes, as it is likely that the complexity of departure time optimization is a function of the number of vehicles, and the number of candidate trip sequences for a vehicle depends on its time left in operation and the number of trip variations (i.e. ordered set of stops defining a trip, e.g. long and short variations in each direction) available. For example, with 20 vehicles and 5 candidate trip sequences per vehicle, there are 5^{20} candidate combinations of trip sequences to evaluate, for each of which departure times must be optimized. The complexity of operations planning problems has led researchers to develop heuristics to optimize transit plans offline, as described in Section 4.2. In real-time applications it is critical to optimize quickly. The complex interdependence between vehicle trajectories and the nonlinearities in the cost function, running times, and demand make plan evaluation and optimization computationally expensive.

Given that the full problem is not tractable, a simplified approach must be adopted. It would be challenging to make progress without first reducing the dimensionality of the problem to attain non-combinatorial complexity. Doing so drastically reduces the solution search space, which makes the problem tractable but sacrifices potentially good solutions. This is a critical aspect of schedule-free operations: potential performance benefits derived from increased flexibility and utilization of real-time information may not be realized without a successful optimization approach, and this success largely depends on how dimensionality is reduced. Good approaches should sufficiently reduce dimensionality while leaving a search space containing good solutions.

A natural approach toward reducing dimensionality is decomposing the full problem into subproblems, one per vehicle, solved sequentially, given sequences for all other vehicles. The complexity of this approach is

$$\mathcal{O}\left(\sum_{v\in V}\sum_{i=1}^{|S_v|} k_i\right) \tag{4.16}$$

which, assuming a constant n number of candidate trip sequences per vehicle and a constant complexity k for departure time optimization (again, only for illustrative purposes), simplifies to

$$\mathcal{O}\left(\sum_{v\in V}\sum_{i=1}^{n} k\right) = \mathcal{O}\left(\sum_{v\in V} kn\right) = \mathcal{O}\left(|V|kn\right) \tag{4.17}$$

For example, with 20 vehicles and 5 candidate trip sequences per vehicle, the number of combinations of trip sequences decreases from 5^{20} to $5 \cdot 20$.

This decomposition approach makes the problem tractable, but drastically reduces the solution search space. Since trip sequences are optimized one vehicle at a time, assumed sequences for the rest of the vehicles affect the costs (and optimality) of each candidate sequence for vehicle v. This makes the order in which subproblems are solved matter. For instance, consider two vehicles, v_1 and v_2, each with three possible trip sequences: A, B, and C for v_1, and D, E, and F for v_2. There are $3^2 = 9$ trip sequence combinations. Suppose that the (unknown) optimal solution selects sequences C and F. Since sequences are optimized separately for each vehicle, a sequence must be assumed for v_2 when optimizing for v_1. Suppose that sequence D is assumed for v_2. Based on this assumption, sequence A might be optimal for v_1. Then, given that choice, sequence D might be optimal for v_2. The selected sequences A and D are suboptimal. The global optimum was not found because the decomposition approach pruned the combination C, F from the search space.

While (4.16) and (4.17) assume a single pass through vehicles (in some order), it might be desirable to do multiple passes to widen the search space (i.e. not reduce it as much) and increase the potential of finding good solutions. In general, if g passes are performed, the complexity is

$$\mathcal{O}\left(g\sum_{v\in V}\sum_{i=1}^{|S_v|} k_i\right) \tag{4.18}$$

which, assuming a constant n number of candidate trip sequences per vehicle and a constant complexity k for departure time optimization, simplifies to

$$\mathcal{O}\left(g|V|kn\right) \tag{4.19}$$

Multiple passes could be done in sequence and in parallel. For example, multiple sequential passes through the single-vehicle subproblems (in some order) would allow revisiting a trip sequence given trip sequences optimized in the previous pass. This could be repeated until there are no further changes in trip sequences, optionally up to a predetermined number of times. Following the previous example of two vehicles, each with three candidate sequences, suppose that when optimizing sequences for v_2 given sequence A for v_1, sequence F is selected (rather than D) for v_2. Since the assumed sequence for

v_2 changed, it is possible that a local optimum has not yet been found, so the sequence for v_1 could be revisited in a second pass. Sequences C and F would be selected for v_1 and v_2, respectively, in this second pass. A third pass would lead to no changes because a locally optimal solution (in this case also globally optimal) has been found. Passes in different orders or with different initially assumed sequences could be processed in parallel, optionally performing multiple sequential passes for each order. For example, three parallel instances of the problem could be solved, each assuming a different sequence for v_2 when optimizing sequences for v_1. After each instance optimizes sequences for v_2, the solution with least cost would be selected. Although the multiple-passes approach improves the probability of finding good solutions, it may not be feasible to guarantee global optimality in problems of real size.

The number of feasible trip sequences increases with time remaining in operation, $u'_v - t_0$, and number of trip variations. Vehicles exiting sooner can serve fewer additional trips. For instance, when $u'_v - t_0 = 15$ minutes, vehicle v may only have time to finish the current trip, so there is a single feasible sequence. In contrast, a vehicle with 4 hours remaining in operation may have time for, say, 4 trips between terminals. Short variations can greatly increase the number of feasible trip sequences. The availability of a single short-turning point per direction can lead to thousands of feasible trip combinations for vehicles having more than 4 hours remaining in operation.

The complexity of departure time optimization must be considered, because a computationally expensive approach could make it infeasible to consider even a single combination of trip sequences. There are several ways to simplify this problem. One is using simple models that can be solved quickly but do not capture all the complexities. For example, the effect of dwell times on running times, the interdependence of vehicle trajectories, and the time-dependence of running times and demand could be neglected, and a linear or convex quadratic formulation could be used. Solving the departure time problem fast allows testing a larger number of feasible sequences, but the stronger assumptions may lead to unrealistic cost estimates, and thus poor choices. A different approach is using richer models with optimization algorithms requiring few objective function evaluations, trading the ability to find globally optimal solutions (of a possibly unrealistic model) for the ability to capture information such as dynamic running times and demand.

It is also possible to combine simple and complex models with hopes of reaping benefits from each. This can be done in stages when the whole problem is decomposed into smaller problems, or at once by employing meta-models mapping values of one model to the other (Osorio and Bierlaire, 2013). A policy approximation approach can be used instead of, or in combination with, departure time optimization. For example, Chapter 2 shows that the even headway heuristic performs well in most cases. This heuristic could be applied within the model to set departure times, subject to constraints on holding and exit times. The result could be used directly to compute the cost of a combination of trip sequences, or as an initial point given to an optimization algorithm for further improvement. Chapter 2 shows that optimization adds value principally when capacities are reached and passengers are left behind, so optimization could be used only when overcrowding is detected in the current or future system states.

Depending on how quickly departure times can be optimized for each trip combination, it may be necessary to discard some sequences without evaluation. For example, sequences in which a vehicle exits with plenty of time to serve more trips may be disregarded. Sequence elimination heuristics of this nature can help reduce the number of required departure time optimizations from thousands to hundreds or tens. Alternative sequences for a given vehicle can be evaluated in parallel.

In the context of this research it is desirable to capture the dynamics of running times and demand, because neglecting them can lead to significant differences between modeled and real costs. For instance, neglecting an increase of running times during peak operations on a transit corridor might cause a simple model to suggest a plan in which drivers are expected to exit on time, but in reality there will be significant exit lateness.

It might be possible to implement schedule-free operations without optimization using heuristics for all the decisions. For example, holding times could be based on target headways from the service plan, and short-turning could be employed when the number of trips between terminals that a vehicle is predicted to serve decreases as a result of delays. In this case, short-turning would be used to correct for the delay and allow the vehicle to serve the original number of trips. With more sophisticated heuristics it might be possible to short-turn in response to overcrowding in addition to delays. A potential drawback of an all-heuristic approach is poor management of consecutive vehicles being delayed. For example, three consecutive vehicles may be delayed enough to qualify for short-turning according to the heuristic, but short-turning them all could lead to very long headways on a portion of the line, and capacity problems could ensue.

4.6 Simplified Methodology

This section presents a specific methodology developed based on the preceding discussion, with the aim of making the schedule-free paradigm operational in a simulated transit line, as presented in Section 4.7. Several simplifications are made in the interest of tractability. Application results presented in Section 4.7.2 suggest that this methodology does not perform well enough and that refinement is needed. Nevertheless, it lays the groundwork for further exploratory work.

Figure 4-4 shows an activity diagram of the optimization process. The process begins by generating a *basic trip sequence* for each vehicle, which becomes the initially assumed trip sequence. Basic trip sequences have vehicles serve complete trips (without stop-skipping) and return to the exit location.

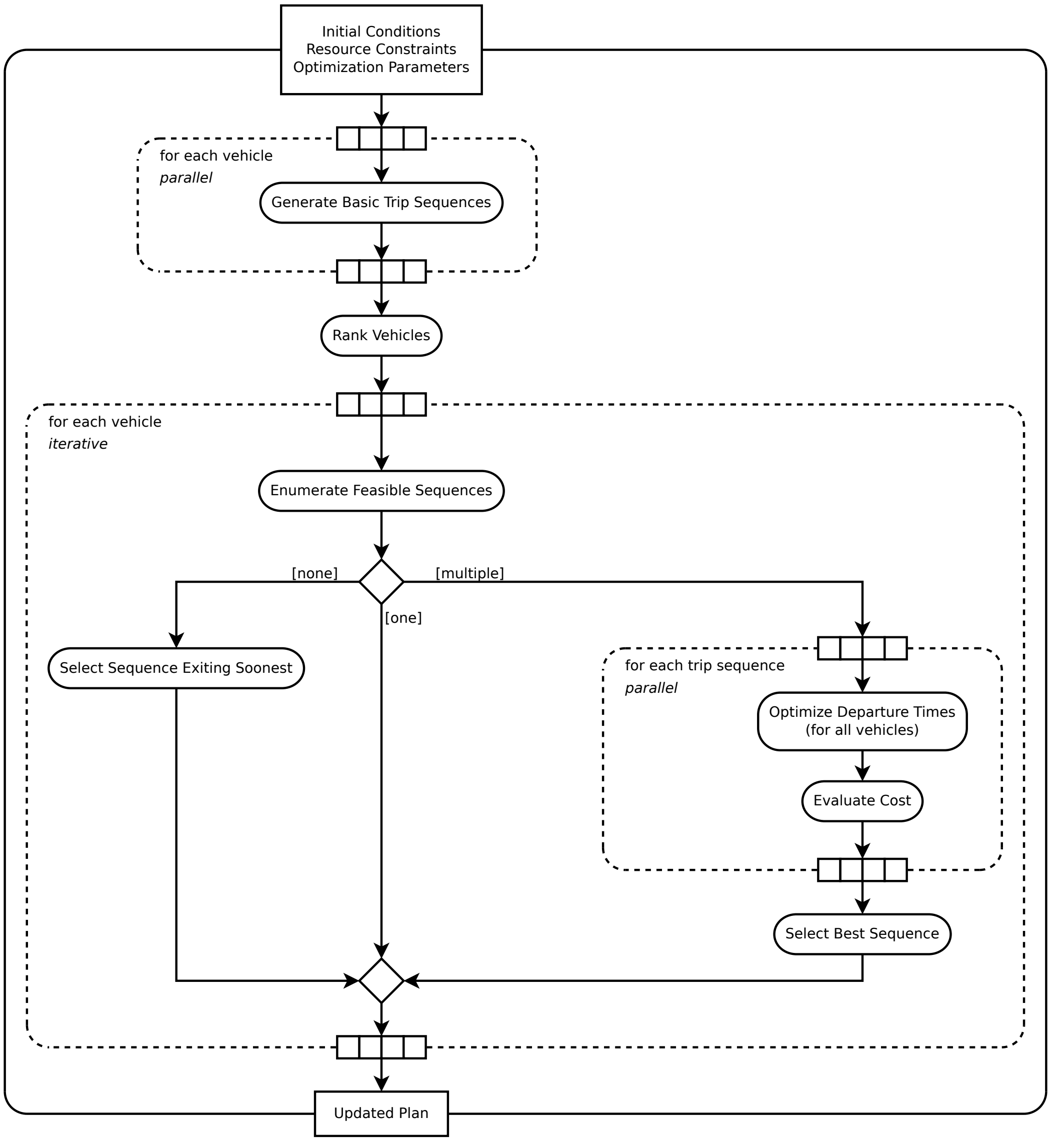


Figure 4-4: Plan Optimization Algorithm

A basic trip sequence is *feasible* if the vehicle exits on, or before, the latest allowed exit time. Trajectories for each vehicle are then optimized, one vehicle at a time, following a specific order (discussed later in this section). Optimized trajectories replace initially assumed ones, such that trajectory optimizations for subsequent vehicles incrementally reflect these updates. Each vehicle's trajectory is optimized by enumerating feasible trip sequences and selecting the one with lower cost. Departure times are optimized as part of each sequence's evaluation. (Following the Unified Modeling Language specification, the keywords *parallel* and *iterative* characterize each *expansion region* denoted by a dashed rounded box. *Parallel* indicates independent activities that can be processed either in parallel or sequentially, with order not mattering, while *iterative* indicates activities that must be processed in sequence, following a certain order if one is established. In this case, the vehicles expansion region processes vehicles in order of ranking, as described later.) The remainder of this section discusses each of these steps in greater detail.

Following the discussion in Section 4.5.1, the trip sequence problem (4.11) is decomposed into sequential trip sequence subproblems for each vehicle, as follows:

$$\underset{s_v \in S_v}{\text{minimize}} \quad C(s_v; s_{\bar{v}}, d_s^*, p) \tag{4.20}$$

where s_v and S_v denote a candidate trip sequence and the set of feasible trip sequences for vehicle v, respectively, $s_{\bar{v}}$ denotes the trip sequences assumed for all other vehicles, and d_s^* denotes optimal departure times for each trip sequence combination (provided by the departure time subproblem). Each instance of problem (4.20) optimizes the trip sequence of a vehicle v and departure times for all vehicles (through the departure time subproblem (4.12)), under assumed exogenous trip sequences $s_{\bar{v}}$ for other vehicles. Starting with an initial assumption about the trip sequences for all vehicles, the subproblems are solved sequentially, once per vehicle, each time capturing previously optimized trip sequences. Thus, when the subproblem is solved for the last vehicle, all trip sequences and departure times have been optimized.

Currently planned trip sequences (from the previous plan update, before t_0) may be assumed for all vehicles to start. Otherwise, a *basic trip sequence* s_v^* may be assumed. Basic trip sequences finish the current trip (if one is being served) as originally planned and have no stop-skipping after the start of the next planned trip. They can be based on one of two approaches.

The first approach generates the *longest feasible basic sequence*, composed of trips between terminals until the latest possible on-time exit, by solving the following problem:

$$s_v^* = \underset{s_v \in S'_v}{\text{argmax}} \quad u_v \tag{4.21}$$

$$\text{subject to} \quad u_v \leq u'_v \tag{4.22}$$

$$h_e = h_e^{\min} \quad \forall e \in E_v \tag{4.23}$$

where S'_v is the set of all basic trip sequences for vehicle v, u_v and u'_v are the planned and target exit times of vehicle v, respectively, and h_e and $h_e^{\min}$ are the planned and minimum allowed holding times for planned stop visits E_v, respectively. Minimum holding times typically imply holding only for the minimum required driver rest time between trips at terminals and turning points. Problem (4.21) is easily solved by adding trips between terminals until the first late exit (after the target exit time u'_v), and then removing the last cycle.

The second approach generates the *closest basic sequence*, composed of trips between terminals until the exit closest to the target exit time, by solving the following problem:

$$s_v^* = \underset{s_v \in S'_v}{\operatorname{argmin}} \quad |u_v - u'_v| \tag{4.24}$$

$$\text{subject to} \quad h_e = h_e^{\min} \quad \forall e \in E_v \tag{4.25}$$

If the closest basic sequence exits early, it is feasible and equal to the longest feasible basic sequence. Infeasible basic sequences must be replaced during the optimization process with sequences that, by serving fewer or shorter trips, satisfy exit constraints. While (4.21) leads to initial assumptions having no exit lateness cost, (4.24) assumes more aggressive usage of vehicles, and may more closely match the (later) optimized trip sequence, which may skip stops to prevent exit lateness. The latter approach is used for the application in Section 4.7 and is assumed in the remainder of this section. Problem (4.24) is easily solved by adding trips between terminals until the first late exit (after the target exit time u'_v), and then removing the last cycle if the previous time at the exit location is closer to the target exit time. The constraint $u_v \leq u''_v$ can be added as a way of allowing closest basic sequences exiting late only if exit lateness does not exceed the hard constraint.

Once each vehicle has an initial trip sequence, departure times can be optimized by solving (4.12) to finish defining an initial solution. Departure time optimization is discussed in Section 4.6.4.

Since trip sequences of each vehicle are optimized separately (though not independently), the order in which vehicles are processed can affect the outcome. Vehicles for which optimal trip sequences significantly differ from initially assumed trip sequences should be processed first. It is useful to consider both the importance and urgency of (potentially) changing each vehicle's basic trip sequence. It is more important for vehicles with exit lateness in their closest basic trip sequence, because alternative trip sequences are required to achieve a feasible plan. For example, a vehicle may need 10 more minutes to complete the last full round trip without exiting late. A single short-turn may allow the vehicle to serve an additional (short) round trip without exiting late, and this could benefit passengers. Replacing initially assumed infeasible sequences with feasible sequences first allows later optimization decisions to reflect these changes. It is more urgent for vehicles exiting sooner, because there may be fewer trip sequences that work well. For example, there may be only one sequence that prevents a vehicle with 50 minutes left before its target exit time from exiting late. By updating trip sequences for the most critical vehicles first, trip sequences for vehicles with greater flexibility can be optimized to work well with those of vehicles having fewer options. Importance and urgency criteria can be combined to rank vehicles as follows:

$$\rho_v = -\max\left(0, u_v^* - u_v'\right) + \theta_T \left(u_v' - t_0\right) \tag{4.26}$$

where u_v^* denotes the planned exit time of vehicle v with the basic trip sequence s_v^* exiting closest to target, and θ_T is a weight establishing the trade-off between importance and urgency.

The first term in (4.26) gives the negative exit lateness, zero if the vehicle does not exit late. The second term gives the time remaining before the target exit. A lower rank ρ_v is given to vehicles with higher potential benefit and urgency. Trip sequences are optimized in increasing order. It is possible to consider more than one ranking function, solving the problem for different vehicle orders in parallel, and selecting the minimum cost plan at the end.

The relationship between candidate sequences is difficult to model because of changes in vehicle order, optimal departure times, interaction between vehicles, and lateness. This makes it challenging, for example, to develop tight bounding rules for a branch and bound algorithm. Instead of attempting this, the best sequence is picked through enumeration. The value of a trip sequence is highly dependent on departure times (of all vehicles from all terminals, turning points, and stops), which affect headways, waiting times, and exit lateness. For each sequence s_v being considered, the departure time optimization subproblem (4.12) is solved. Feasibility is determined by exit lateness constraints and minimum and maximum allowed holding times. Departure times are manipulated through holding at stops and terminals. Although the optimization problem is equivalent to the holding control problem, it is departure times rather than holding times that define a plan (along with trip sequences). Exit lateness cost is incurred when $u_w' < u_w \leq u_w''$. Lateness cost of vehicles with infeasible basic sequences can be neglected until their sequences are replaced with feasible ones.

Computation times for plan optimization depend on the complexity of the departure time optimization algorithm. There is a trade-off between sophistication and speed. A general purpose nonlinear optimization algorithm can be used to minimize cost, but if this takes too long then it may not be possible to evaluate all candidate trip sequences in the required time. Chapter 2 shows that (except in cases of overcrowding) holding optimization generally results in even headways. It is therefore reasonable to approximate the departure time optimization policy with an even headway algorithm requiring far fewer performance model evaluations, in exchange for the ability to evaluate more candidate trip sequences. Since candidate plans for a vehicle are mutually independent, they may be evaluated in parallel. The departure time subproblem could be solved fully for all feasible trip sequences with distributed computing.

The methods employed are deterministic. Aside from the implications of neglecting stochasticity on the optimality of operations plans generated by this approach, this can lead to unplanned late exits because trips can take longer than expected. In effect, the exit lateness policy prevents (or discourages) trip plans with expected late exits (at the time of trip plan optimization), rather than late exits per se. While some operators might accept this, other may require a stricter policy. It is possible to establish upper bounds on exit times based on high-percentile running times to obtain lateness estimates under pessimistic (i.e. slow run) scenarios.

Key components of the real-time planning optimization process include (1) a discrete event performance model that captures dynamic running times and demand, the interdependence of demand and running times through dwell times, mean passenger cost including waiting time at stops and in-vehicle time, and driver lateness, (2) a simple running time model that captures the time-dependence of running times but neglects the endogenous contribution of dwell times, (3) a trip sequence optimization algorithm, and (4) a departure time optimization algorithm. These are discussed in the following sections.

4.6.1 Event-Based Performance Model

The performance model makes deterministic predictions of system evolution over a fixed duration time horizon. Inputs defining the initial state include vehicle arrival times and locations, most recent departure times from each stop, and number of passengers in vehicles and at stops. Demand is modeled by origin-destination pair, using continuous variables as in Chapter 2. Running times and demand are modeled as functions of time. It is assumed that vehicles stop at all stops. Inputs defining planning and control decisions include the planned sequence of stops to be visited by each vehicle and planned departure times by vehicle and stop. Other inputs include boarding and alighting time per passenger, horizon length, minimum departure headways, vehicle capacities, minimum and maximum holding times at terminals, turning points, and stops, and weights for waiting time, in-vehicle time, and driver lateness. Outputs include vehicle arrival and departure times, number of boardings and alightings at each stop, loads, and cost.

The performance model is based on events representing vehicle arrivals at stops. Events are processed chronologically, allowing vehicle order to change over the prediction horizon. Changes in vehicle order may be caused by vehicles entering or exiting service, short-turns, and overtaking. An event heap is initialized to contain events representing the first stop visit of each vehicle, including vehicles that have not yet entered service. Each stop visit is processed as follows:

1. The number of alightings, total alighting time, and remaining vehicle capacity are determined.

2. If the vehicle is at its last planned stop, the vehicle is taken out of service. No new events are generated.

3. The number of boardings, total boarding time, and number of passengers left behind (when vehicles reach capacity) are determined. The number of passengers who wish to board includes passengers waiting in the initial state, passengers left behind by previous vehicles, and passengers arriving during the prediction horizon. Passenger arrivals are determined by integrating demand functions, which give arrival rates over time. We assume that passengers only board vehicles serving their destination in the current trip.

4. The departure time is determined, considering total alighting and boarding times, planned departure time, holding time constraints, and minimum departure headway. Before considering constraints, a vehicle holds until its planned departure time, or departs immediately if it is late. Holding constraints may require a vehicle to hold for a minimum duration (e.g. at a terminal) even if it is late, or may limit how long a vehicle can hold, forcing it to depart before the planned time. Minimum departure headways may extend a vehicle's holding time, even if the planned departure time has passed.

5. The running time to the next stop is determined by evaluating the running time function for the link connecting the current stop and the next stop. The arrival time at the next stop is determined by adding this running time to the departure time from the current stop. A new event is created representing the next stop visit, and added to the heap.

4.6.2 Exogenous Trip Running Time Model

Estimating the time it takes a vehicle to complete a trip is a critical part of the schedule-free plan optimization framework. A vehicle might exit early or late depending on its planned trip sequence and the duration of each trip. The discrete event performance model provides the space-time trajectories of each vehicle, from which exit time and exit lateness are derived, but it requires a plan (stop sequences and departure times) as input and is computationally expensive to evaluate. In order to generate candidate trip sequences and evaluate their feasibility with respect to exit lateness, many trip durations must be estimated; vehicle movement at the stop level, loads, and passenger cost are not necessary at first. The role of the exogenous trip running time model is to provide these running time estimates efficiently. The model consists simply of linear piecewise functions, one per variation, mapping time of day to trip running time. If not supplied as an input, trip times may be obtained from the output of the discrete event performance model. The resulting univariate trip running time model neglects the effect of demand and the interdependence of vehicle trajectories, treating them as exogenous factors related to time, but it is precise enough to estimate the exit lateness resulting from candidate trip sequences.

4.6.3 Trip Sequence Optimization

The schedule-free optimization process determines which trip sequence each vehicle should serve in order to minimize cost while satisfying exit constraints. The availability of multiple trip sequences enables efficient utilization of vehicles and crew while satisfying exit time and location constraints. If restricted to serving trips between terminals of a single line, vehicles may reach their exit location having enough time to serve shorter trips, but not enough to run another full cycle, and must therefore exit early to satisfy exit constraints, wasting the time left between the end of the last trip and the target exit time. Stop-skipping strategies such as short-turning, dead-heading, expressing, and limited stop service offer the possibility of serving shorter trips and reducing waste. In addition, there may be the possibility of serving trips on a set of lines sharing a terminal. If the different lines have cycles of different durations, interlining could help increase the efficiency of deployed resources. Besides preventing waste of resources, trip sequence optimization may help manage overcrowding with targeted capacity increases.

The trip sequence optimization model chooses trips from a set of variations. A *variation* is a unique ordered set of stops beginning and ending at a turning point. In this context, a *turning point* is a stop (with or without demand) where trips can begin or end. We assume that vehicles begin all trips empty and that passengers only board vehicles that will serve their destination in their current trip. Figure 4-5 depicts a transit line with four turning points: A and D at the terminals and B and C en-route. In this case turning points B and C are, like terminals, (dummy) stops without demand where a vehicle can hold between trips. Variations AD and DA run from terminal to terminal, while variations AC, CA, DB, BD, BC, and CB are short trips. Short-turns are enabled by specifying variations beginning or ending at en-route turning points. A vehicle at a turning point can be taken out of service (if the turning point is the designated exit location and there is no time left to serve more trips) or continue to serve trips on any of the variations starting at that turning point. For example, a vehicle at A might offer the next trip on variations AC (short) or AD (long). Dead-heading is enabled by specifying *dead variations*, which begin and end at turning points but have no stops in between. For instance, a dead variation starting at A and ending at B enables a vehicle currently at A to deadhead to B and begin revenue service on variation BD. Limited stop services are enabled by specifying variations that skip stops. The problem's complexity increases exponentially with the number of variations.

Expressing and short-turning trips can force passengers to alight. A benign form of these strategies is considered in this research: short-turning and expressing must be decided (and announced) at terminals or turning points between trips, so that passengers are not forced to alight as a result of stop-

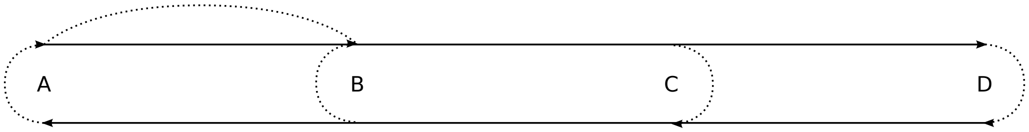


Figure 4-5: Schematic of Transit Line with Turning Points

skipping before completing their trips. The unrestricted versions of these strategies, not considered here, are of interest for responding to major incidents. These strategies could be enabled by capturing the inconvenience imposed on passengers in the objective function. The following discussion assumes that vehicles must finish their current trip. For example, if a vehicle is currently running a trip on variation AD, the real-time trip sequence optimization algorithm may not change the current trip to AC, because that would force passengers whose destinations are downstream of C to alight to complete their trip. However, the algorithm can choose the sequence of trips to be offered once the vehicle arrives at D. Assuming that passengers only board vehicles that will serve their destination during their current trip, the algorithm does not force alightings, and capturing the corresponding cost becomes unnecessary.

Trip sequence optimization begins by generating a basic trip sequence for each vehicle, given by (4.24), or obtained from the previous real-time plan (to be replaced by the one being created). Minimum holding times are assumed, as described earlier. This means that stop visit times are based on running times between stops, without holding to regulate headways except where required. For example, vehicles may be required to hold at least 2 minutes between trips. A set of candidate feasible trip sequences is generated for each vehicle, as shown in Figure 4-6. Time is shown on the vertical axis for a vehicle currently on its way to A, having a target exit at terminal A at time u', with a hard exit constraint at time u''. Feasible trip sequences finish the current trip, do not have trips beginning after u', and exit before u''. The exogenous trip running time model is used. Hence, the set of feasible trip sequences is $\{A\}$, $\{A, D, A\}$, $\{A, C, A\}$, and $\{A, C, B, C, A\}$. The latter has some exit lateness cost. Sequences $\{A, D, B, C, A\}$, $\{A, C, A, C, A\}$, and $\{A, C, B, D, A\}$ are infeasible because they exit after u''. Sequence $\{A\}$ finishes with the current trip after which the vehicle is taken out of service.

If there are no feasible trip sequences, the trip sequence that returns the vehicle to its exit location soonest is selected. If there is a single feasible trip sequence, it is selected. If there are multiple feasible trip sequences, each one is evaluated by solving the departure time subproblem (4.12), and the one resulting in the least cost is selected, thus solving (4.20). Trip sequences ending after the target exit time u' incur exit lateness cost.

A number of measures are taken in the interest of tractability. When a vehicle has time to serve several more trips before exiting, there can be

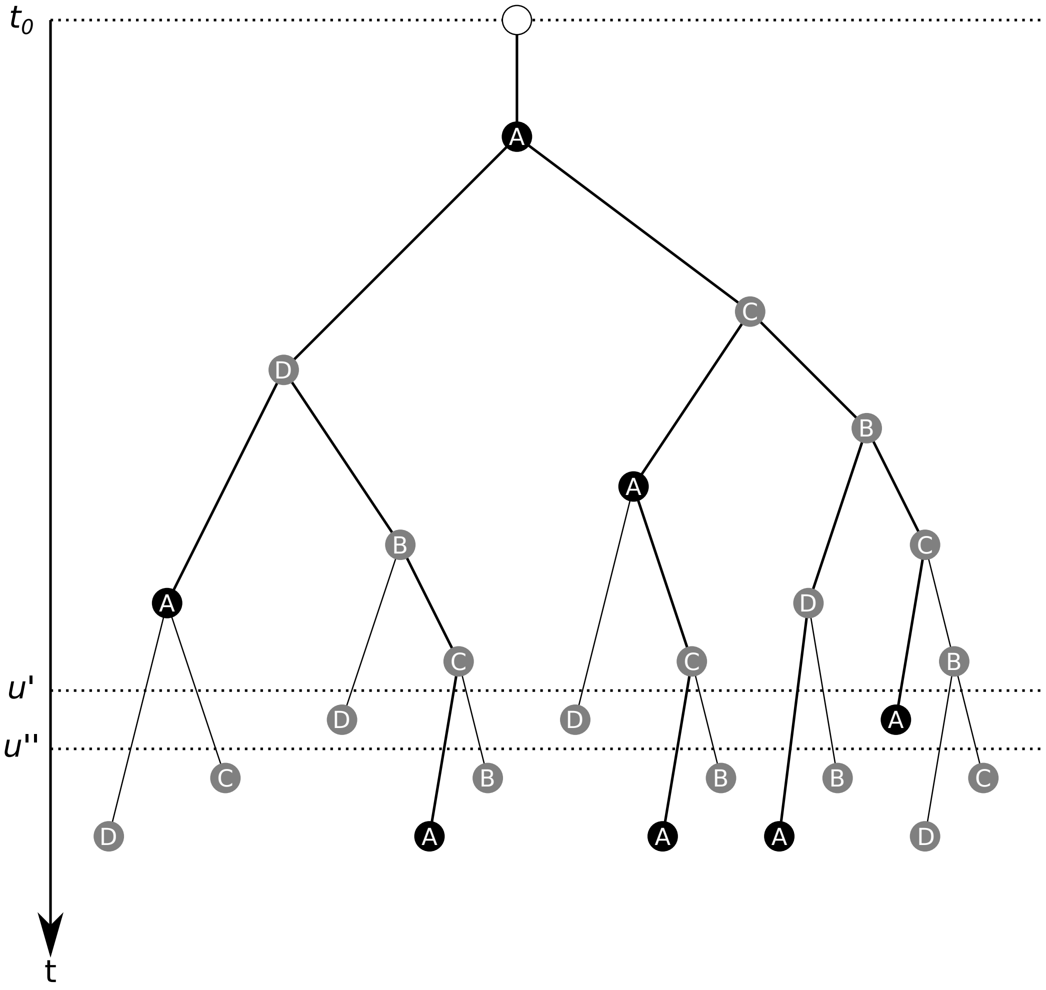


Figure 4-6: Trip Sequence Optimization for a Single Vehicle

thousands of feasible sequences. The following heuristics can be applied to speed up trip sequence optimization:

1. Optimization can be skipped for vehicles scheduled to enter the system after the end of the optimization horizon, since their trajectories should have little or no effect on the optimal plan for vehicles in the horizon.

2. Trip sequences having excessively early exits, with time to serve more trips and not exit late, can be excluded. In particular, parents of feasible trip sequences, such as sequence $\{A\}$ in Figure 4-6, can be excluded. Consider a set of feasible trip sequences T_v for vehicle v. A sequence $p \in T_v$ is a parent if and only if there is another sequence $q \in T_v$ such that $p \subset q$ and $|p| < |q|$.

3. If the basic sequence s_v^* given by (4.24) is feasible, it can be selected without optimization, especially when the intent is for vehicles to serve full trips and the operator prefers not to employ stop-skipping strategies unnecessarily. From the three feasible sequences of the previous example, $\{A, D, A\}$ would be selected based on this criterion.

4. If s_v^* is infeasible, then either one less full trip can be served or stop-skipping strategies must be employed, but the operator might prefer not to consider trip sequences having more stop-skipping than necessary. If so, trip sequences skipping stops after lateness $u_v^* - u_v'$ has been absorbed through stop-skipping strategies can be excluded. For example, for a vehicle with 5 minutes of exit lateness in its basic sequence, feasible sequences with a single short-turn that saves at least 5 minutes would be retained, but those having additional short-turning would be excluded.

Excessive elimination of candidate sequences reduces the possibility of finding an optimal plan. On the other hand, eliminating sequences that are unlikely to be optimal speeds up real-time planning and can even improve the solution to the general (all vehicles) problem. Since the planning problem is broken into separate problems for each vehicle, the optimal sequence for a vehicle's subproblem, given sequences assumed or previously optimized for other vehicles, may not be the optimal sequence for the general problem. Retaining sequences that are good globally can improve the solution. These set reduction heuristics are investigated in Section 4.7.3.

4.6.4 Departure Time Optimization

The departure time optimization problem (4.12) seeks a combination of departure times that minimizes cost. While trip sequence optimization relates to the spatial dimension of the operations plan (although minimum holding times are assumed to determine feasibility with respect to exit time constraints), the departure time optimization defines the temporal dimension. Departure times heavily influence headways, waiting times, crowding, and exit lateness. Candidate plans are not fully formed until both spatial and temporal variables are defined, so their cost cannot be quantified before optimizing departure times. Since the discrete event performance model is used to evaluate the objective function, departure time policies capture planned trip sequences. The horizon should cover the latest exit of all vehicles for which plans are being optimized.

The starting point for departure time optimization is the set of stop visits planned by the trip sequence optimizer, with minimum holding times. Departure time optimization can add holding time beyond the required minimum, potentially decreasing cost to passengers but also increasing exit lateness. Exit lateness is discouraged through the exit lateness component C_L in objective function (4.7), which allows trading off acceptable amounts of lateness for improved service quality, while the hard constraint on exit lateness can prevent additional holding regardless of its benefit to passengers.

The departure time problem is related to holding time optimization and headway optimization, since holding times influence departure times, which in turn influence headways. Manipulating headways would be the most direct way to affect cost, but it would require calculating departure times for all vehicles simultaneously, since the departure time of the first vehicle affects the departure time of all following vehicles given headways. In addition, it is possible that the holding times required to achieve specific headways violate holding time constraints. For example, 3 minutes of holding at a stop may be needed to achieve a 5 minute headway, but this would be infeasible if vehicles can hold at most 1 minute at that stop. Manipulating holding times is the most indirect way of affecting cost, because it requires calculating departure times and then headways. Since vehicle trajectories affect trajectories of upstream vehicles, changing the holding time of a vehicle can affect the departure time, and optimal holding time, of a nearby upstream vehicle. Manipulating departure times requires verifying that the holding times required to achieve specific departure times satisfy lower and upper bounds, but planned departure times can be set, and the corresponding holding times verified, independently for each vehicle, which can improve the performance of some optimization algorithms. When a vehicle arrives at a stop after the planned departure time, it leaves as soon as possible. When the holding time required to depart as planned exceeds the allowed maximum, the vehicle holds for the maximum allowed time and leaves before the planned departure time.

The departure time problem is nonlinear and (in general) non-convex, so there may be multiple local minima. The feasible solution space can be very large for problems of typical size, and grows super-linearly with the number of planned events at control points in the horizon. Dimensionality can be reduced by optimizing departure times at stops only before the next arrival at a turning point, and thereafter only at turning points. Departure times from stops not being optimized are determined based on forward propagation using minimum holding times, dwell times, and running times. Non-convexity makes the quality of solutions found by local optimization algorithms dependent on the initial point. A good initial point can also speed up convergence. A *constrained even headway* policy is applied iteratively to transform the minimum holding solution from the trip sequence optimizer into an initial point with lower cost, ensuring that holding time constraints are satisfied and vehicles do not exit late. Since the even headway policy is insensitive to cost, it can escape local valleys in the objective function. The policy approximation approach may be sufficient to obtain good solutions, but a nonlinear optimizer can be used to further reduce cost by fine-tuning planned departure times within a trust region. This can be of value in situations when even headways are thought to be suboptimal, e.g. when vehicles fill and passengers are left behind. Given that the problem is non-convex, this approach does not guarantee global optimality, and metaheuristics such as simulated annealing could be used to improve the solutions found.

An activity diagram of the constrained even headway algorithm is shown in Figure 4-7. The algorithm combines evaluations of the performance model with a value approximation approach for vehicle trajectories. Planned events where holding is allowed are processed in order of departure time. Preceding and following headways are calculated at a reference stop, which can be different from where the vehicle holds. For example, when adjusting the departure time from an en-route turning point, the first downstream stop can be the reference, in order to capture headways as experienced by passengers. After calculating the change in departure time leading to equal preceding and following headways subject to constraints on holding and exit lateness, the trajectory of the vehicle for which departure time is being adjusted is approximated, shifting the downstream trajectory by the change and neglecting effects on other vehicles. Adjusted trajectories represent target departure times. The performance model is evaluated after approximating all planned events in order to calculate cost and capture the interaction between vehicles. The approximation-evaluation process can be repeated several times to increase headway regularity.

4.7 Application

One of the objectives of this research is to assess the potential of the schedule-free paradigm. While the previous sections have discussed the conceptual arguments for planning trips in real-time, it is also important to demonstrate the paradigm's feasibility and performance. To that end, this section discusses the application of the schedule-free paradigm to a simulated high-frequency transit line, described in Section 4.7.1. Feasibility is evaluated in terms of computational cost and, in particular, optimization times. A formulation that takes hours to solve could be interesting for off-line applications but is of little value in a real-time context. Section 4.7.2 compares the performance of the transit line under the schedule-based and schedule-free paradigms in cases of no delays, moderate delays, and severe delays, using the simplified methodology

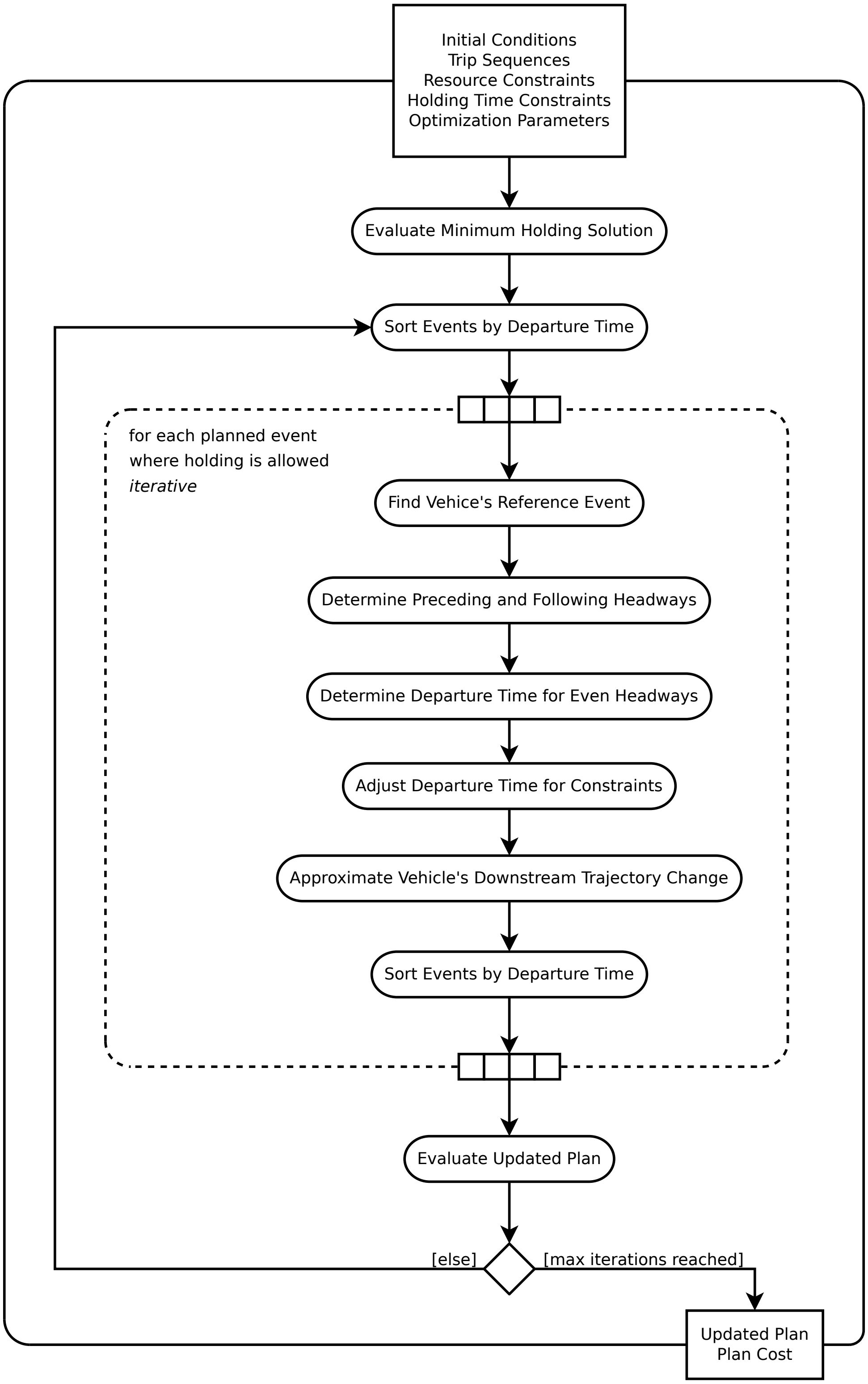

Initial Conditions
Trip Sequences
Resource Constraints
Holding Time Constraints
Optimization Parameters
Evaluate Minimum Holding Solution
Sort Events by Departure Time
for each planned event
where holding is allowed
iterative
Find Vehice's Reference Event
Determine Preceding and Following Headways
Determine Departure Time for Even Headways
Adjust Departure Time for Constraints
Approximate Vehicle's Downstream Trajectory Change
Sort Events by Departure Time
Evaluate Updated Plan
[else]
[max iterations reached]
Updated Plan
Plan Cost

Figure 4-7: Constrained Even Headway Algorithm

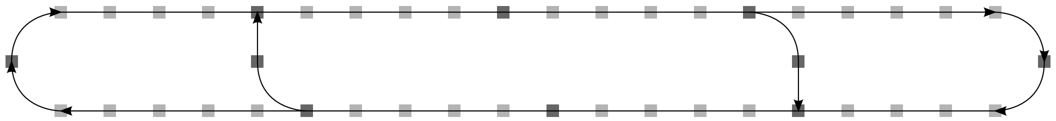


Figure 4-8: Simulated Transit Line

presented in Section 4.6. Performance is evaluated in terms of passenger cost, i.e. waiting times and in-vehicle time, and driver exit lateness. Section 4.7.3 explores the use of heuristics to meet the tractability requirement.

4.7.1 Transit Line

The transit line is a simple (non-branching) transit line with 20 stops per direction and a terminal at each end, as shown in Figure 4-8. Short-turning is allowed at the 15th stop in each direction (to the 5th stop in the opposite direction), but must be decided by the time vehicles start their trips. Vehicles stop at a turning point when short-turning, where they may hold before beginning the next trip. This allows trips running between stops 1 and 20, 1 and 15, 5 and 20, and 5 and 15 in each direction. Deadheading and expressing are not allowed. Terminals and en-route turning points are modeled as (dummy) stops without demand.

There are 25 vehicles (not all operating simultaneously), each with capacity for 60 passengers. Figure 4-9 shows scheduled trips for each vehicle by time; each horizontal line segment indicates a trip. All scheduled trips run between terminals. The schedule was generated using a greedy algorithm that captures running times, demand, and target headways. New trips are dispatched over a period of 8 hours, 95 in each direction.

The running time between stops is (deterministically) 1 minute, except in direction 2 during the peak period between 3:00 and 6:00, when running times increase to 2 minutes per link, to model the typical effect of peak traffic in a shared right of way. We also consider cases in which there are delays in direction 2 during the peak period. Moderate and severe delays cause running times to peak at 3 and 4 minutes per link, respectively, rather than the typical 2 minutes per link. Link running times by direction and time are shown in Figure 4-10. The three running time cases can be considered as different states of the operating environment, reflecting running time variability across days. One can imagine that the line most commonly operates without delays, while moderate delays are encountered occasionally and severe delays rarely, e.g. due to bad weather. The schedule assumes no delays.

Since running times are deterministic (within each case), the ability of the schedule-free paradigm to deal with within-day stochasticity is not observed. The decision to use deterministic running times was made for the sake of simplicity in this first implementation of the new paradigm. With deterministic running times and the same demand across cases (derived from a Poisson pro-

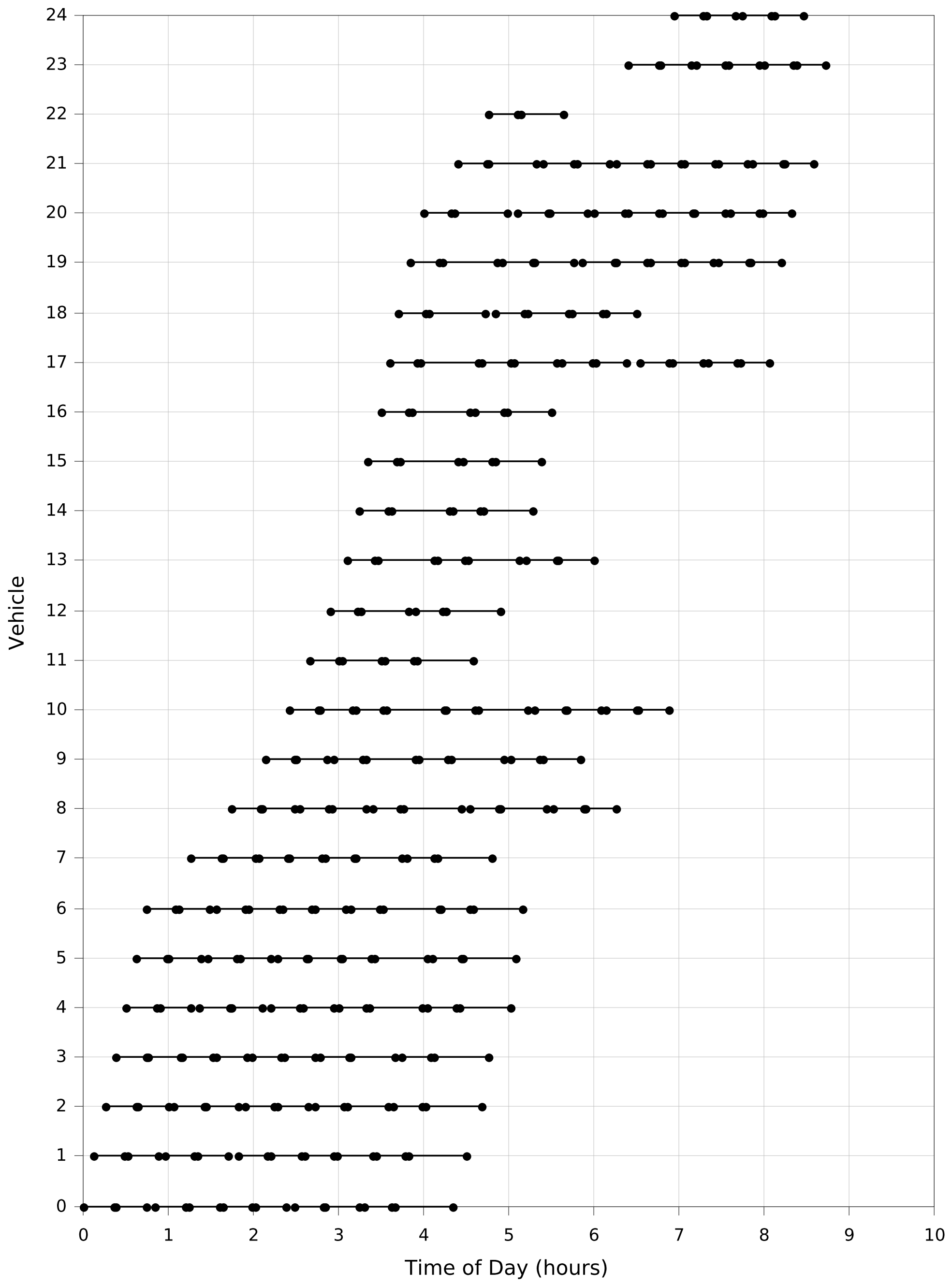


Figure 4-9: Scheduled Duties by Vehicle

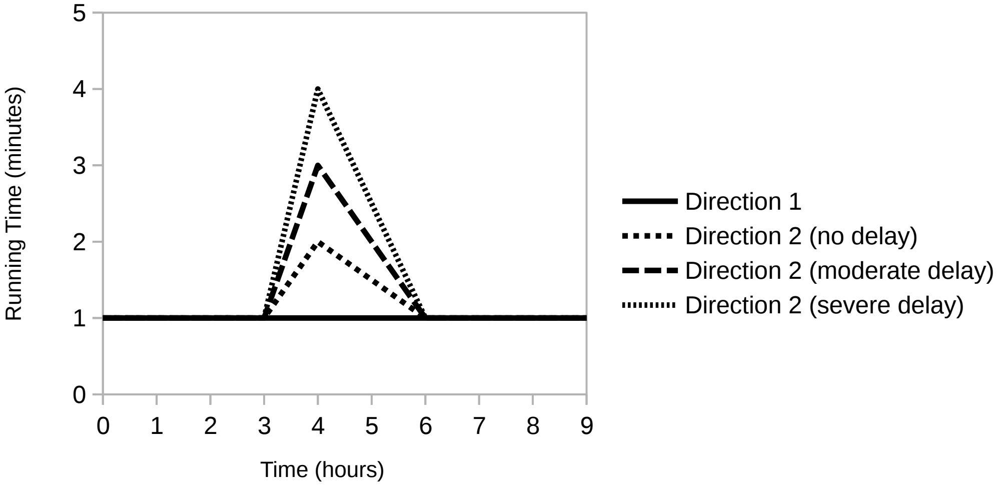


Figure 4-10: Link Running Times

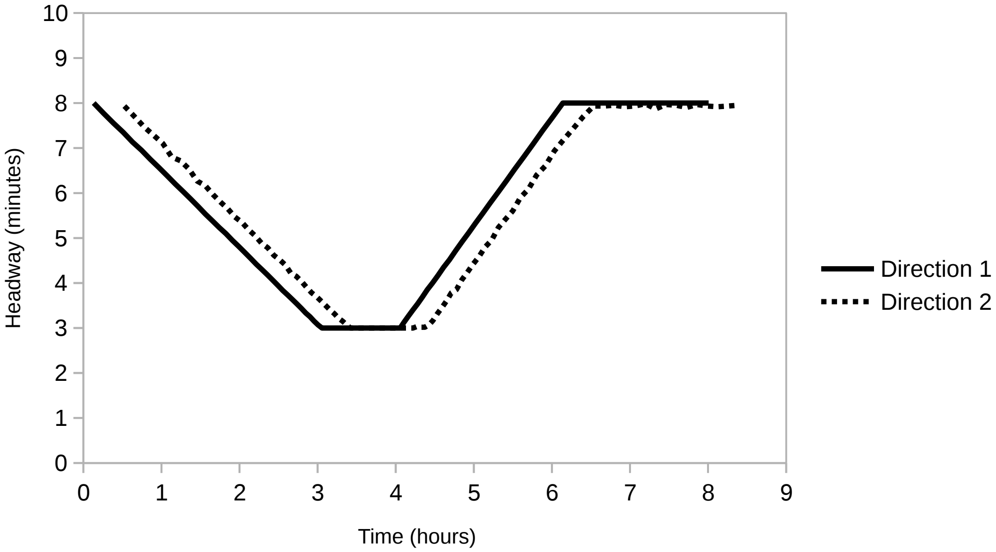


Figure 4-11: Dispatch Headways

cess but with a fixed, common seed for the pseudorandom number generator), the differences in outcomes between cases of the same delay are entirely due to the paradigm difference. (This approach differs from the one followed in Chapter 2, where running time variability was used to generate disruptions and performance measures were calculated across simulation replications for each case.)

The target headway used to generate the schedule is 8 minutes in the off-peak and 3 minutes in the peak, as shown in Figure 4-11. Demand is modeled as a Poisson process. All origin-destination pairs (in each direction) have the same arrival rate function. This form of demand has most boardings at the first stop, most alightings at the last stop, and maximum load in the middle. This pattern may make short-turning less favorable than in transit services having less demand for travel to and from stops close to the ends of the line, because short-turns make vehicles skip the stops with highest arrival rates. The arrival rate λ_1 in direction 1 is such that vehicle loads reach half the capacity when headways are 8 minutes. The arrival rate λ_2 in direction 2 is the same off-peak, but increases such that vehicles are 90% full at the maximum load point when headways are 3 minutes in the peak. A separate pseudorandom number generator is used to generate demand, with common random numbers across all cases.

The target exit time u' is 15 minutes after the end of each vehicle's last scheduled trip, i.e. in Figure 4-9, 15 minutes after each vehicle's rightmost dot. This is also the latest allowed exit time, u''. Since $u'_v = u''_v \quad \forall v \in V$, there is only a hard exit lateness constraint, and it is not relevant to set an exit lateness cost weight (θ_L in (4.7)) or an exponent α in the lateness cost function (4.9). Solution complexity cost is neglected, with $\theta_C = 0$ in (4.7). When evaluating passenger cost, waiting time at the stop is considered twice as onerous as in-vehicle time, with waiting time weight $\theta_W = 2$. The cost discount factor is set so as to halve costs every hour. (These optimization parameters are described in Section 4.5.) Vehicles must hold at least 2 minutes at terminals and en-route turning points (which, like terminals, are modeled as stops without demand), and can hold at most 2 minutes at stops 5, 10, and 15 in each direction. Figure 4-8 shows terminals, en-route turning points, and stops where holding is allowed in darker gray.

Under the schedule-based paradigm, vehicles are held at terminals until their scheduled departure time. Vehicles are dispatched to run short only when current lateness exceeds the time savings expected by short-turning, regardless of exit time. Time savings are calculated based on scheduled running times on skipped portions of the route, assuming minimum required holding between trips. For example, a vehicle departing 10 minutes late from a terminal would short-turn only if it is expected to start the next (return) trip on time or late, in spite of the time saved through short-turning. If instead the vehicle is expected to arrive at the turning point early, and have to hold for more than the required minimum holding time to begin the next trip on time, it would be dispatched to serve a complete trip. This policy applies short-turning only when vehicles are significantly delayed.

Under the schedule-free paradigm, trip sequences and departure times are optimized every 5 minutes to update the operations plan, following the methodology presented in Section 4.6, and employing the sequence elimination heuristics listed in Section 4.6.3. Vehicles are held at terminals or en-route turning points until their planned departure time, and they are dispatched to run short when the plan specifies. The real-time plan optimizer assumes no-delay running times when predicting vehicle trajectories in the three cases.

4.7.2 Results

The results presented in this section demonstrate the feasibility of the schedule-free paradigm, and its performance under the simplified methodology presented in Section 4.6.

Table 4.2 compares performance measures for schedule-based (SB) and schedule-free (SF) operations, across the three cases with different delays, with and without short-turning allowed. The reported waiting, excess waiting, and in-vehicle times are means over all passengers at all times. Exit lateness is the time spent in operation after $u' = u''$, i.e. more than 15 minutes after last scheduled stop visit. Distributions of exit lateness are shown in Table 4.3. For example, in the case of schedule-free operations, with short-turning, when the line experiences a moderate delay, 19 vehicles exit on time (or early), 5 vehicles exit no more than 2 minutes late, and 1 vehicle exits more than 4 minutes late.

There is no significant difference in mean waiting times, excess waiting times, in-vehicle times, or lateness between the two paradigms in the base case. This is not a trivial outcome because the real-time planner does not have the schedule as a reference. Short-turning occurs three times in schedule-free operations (in the case it is allowed), all with the same vehicle. It is not required to prevent a late exit, but it is nonetheless planned by the optimizer, which implies it is driven by a lower predicted passenger cost.

Figure 4-12 illustrates the trajectory of this vehicle under schedule-based and schedule-free operations, compared to the scheduled trajectory. Time is shown on the horizontal axis, and space on the vertical axis, with the top and bottom being opposite terminals. A dot indicates the latest allowed exit time, u'', which is 15 minutes after the last scheduled stop time. As expected, the schedule-based observed trajectory closely matches the schedule. By short-turning three times and using the extra 15 minutes, the schedule-free real-time planner manages to serve another cycle.

This aggressive plan, with no recovery time left at the end of the vehicle's run, works in this case because there are no delays, but must be revisited in the other two cases once the vehicle is delayed. Figures 4-13 and 4-14 show the trajectories of the same vehicle in the cases of moderate and severe delays, with short-turning allowed. The first short-turn is planned as before, but a pair of complete trips (i.e. between terminals) is served rather than a second pair of short trips. There is no further short-turning under schedule-free operations in

Table 4.2: Performance Comparison

		Short-Turning		No Short-Turning	
Delay	Performance Measure	SB	SF	SB	SF
None	Waiting Time (min)	2.6	2.6	2.6	2.6
	Excess Waiting Time (min)	0.0	0.0	0.0	0.0
	In-Vehicle Time (min)	9.6	9.5	9.5	9.5
	Late Exits	0	0	0	0
	Max Exit Lateness (min)	0.0	0.0	0.0	0.0
	Trips	190	192	190	190
	Short Turns	0	3	—	—
Moderate	Waiting Time (min)	3.3	2.7	3.4	2.7
	Excess Waiting Time (min)	0.7	0.1	0.8	0.2
	In-Vehicle Time (min)	10.8	10.7	10.7	10.7
	Late Exits	0	6	1	5
	Max Exit Lateness (min)	0.0	4.1	0.6	1.0
	Trips	190	190	190	186
	Short Turns	2	2	—	—
Severe	Waiting Time (min)	6.3	4.1	3.6	3.7
	Excess Waiting Time (min)	3.7	1.5	1.0	1.1
	In-Vehicle Time (min)	11.7	11.8	12.0	11.9
	Late Exits	7	12	15	9
	Max Exit Lateness (min)	9.6	11.3	10.6	10.8
	Trips	190	186	190	162
	Short Turns	16	21	—	—

Table 4.3: Exit Lateness Distributions

			Exit Lateness (minutes)						
	Delay	Planning	0	2	4	6	8	10	12
Short-Turning	None	SB	25	–	–	–	–	–	–
		SF	25	–	–	–	–	–	–
	Moderate	SB	25	–	–	–	–	–	–
		SF	19	5	–	1	–	–	–
	Severe	SB	18	1	–	2	2	2	–
		SF	13	6	1	1	2	1	1
No Short-Turning	None	SB	25	–	–	–	–	–	–
		SF	25	–	–	–	–	–	–
	Moderate	SB	24	1	–	–	–	–	–
		SF	20	5	–	–	–	–	–
	Severe	SB	10	1	1	3	5	4	1
		SF	16	–	2	2	2	2	1

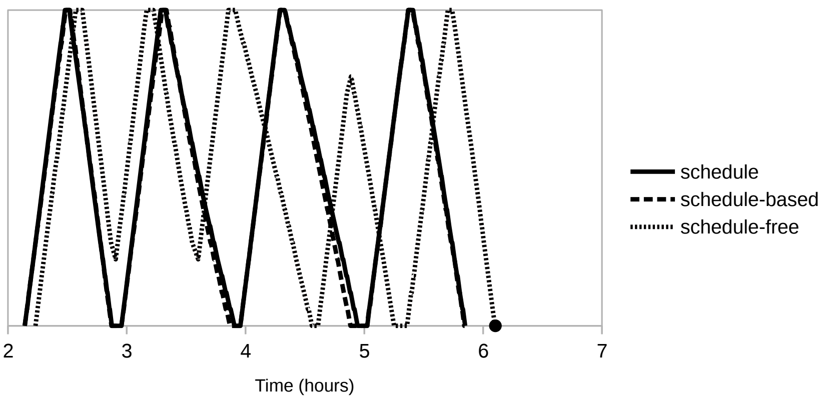


Figure 4-12: Trajectory of Vehicle with Short-Turning, No Delay

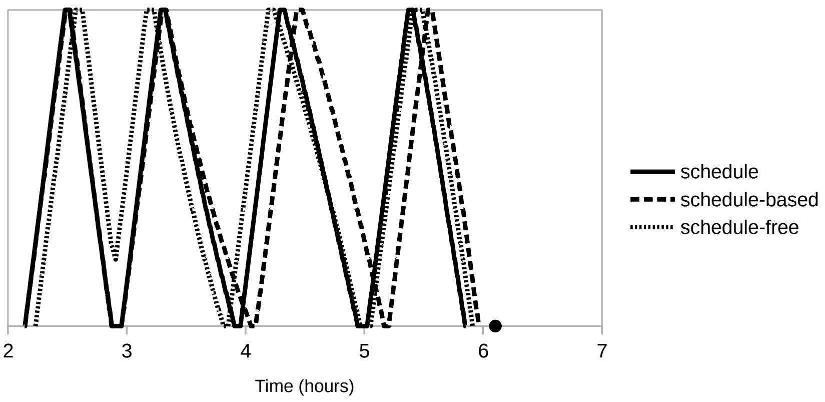


Figure 4-13: Trajectory of Vehicle with Short-Turning, Moderate Delay

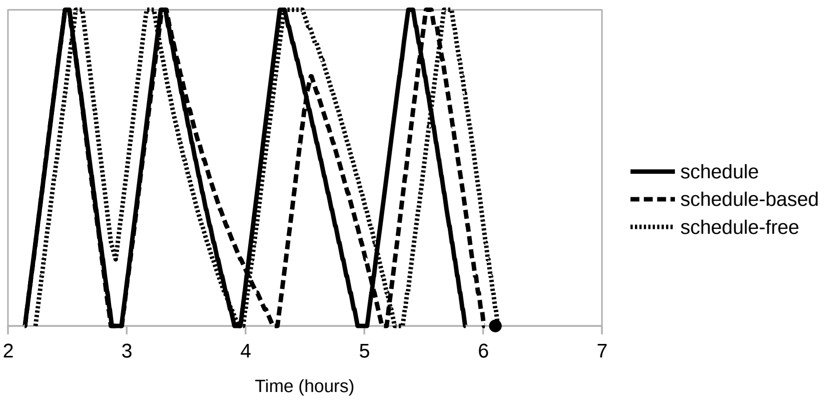


Figure 4-14: Trajectory of Vehicle with Short-Turning, Severe Delay

either of the cases with delays. Despite some recovery time at terminals, this vehicle does not fully recover the schedule (under schedule-based operations), although it does not exit late because of the 15 minutes of recovery time at the end of the run and, in the case of severe delays, one short-turn.

With moderate delays and short-turning allowed, mean waiting time decreases by 0.6 minutes (19%) going from schedule-based (SB) to schedule-free (SF) operations, and mean in-vehicle times differ by less than 0.1 minutes. Excess waiting time decreases by 87%. There are no late exits with schedule-based operations because, aside from two short-turns, lateness is mostly absorbed by the 15 minute grace period after the last scheduled stop visit; although the delay makes some vehicles exit after their last stop's scheduled time, none are delayed by more than 15 minutes, which is when we begin counting exits as late. These results show that at most two short-turns are necessary to prevent exit lateness. Although the schedule-free plan optimizer attempts to prevent vehicles from exiting late (i.e. more than 15 minutes after each vehicle's last stop's scheduled time), 6 vehicles exit late with schedule-free operations, 5 with exit lateness not exceeding 2 minutes, and one 4.1 minutes late. The same number of trips is served with both paradigms, and short-turning is employed twice in both cases.

Vehicles exit late under the schedule-free paradigm because the real-time planner assumes no delays when forecasting vehicle trajectories and determining if they are feasible. Since trips between terminals may not be short-turned once started, lateness cannot always be prevented when a vehicle is delayed in its last cycle. For example, the vehicle that exits 4.1 minutes late enters the system at 2:54 and is supposed to exit by 5:09. Figure 4-15 shows this vehicle's schedule-based, schedule-free, and scheduled space-time trajectories. When the vehicle reaches the exit terminal at 4:04, it has already experienced delays, but the planner assumes no additional delays and, therefore, enough time for the vehicle to serve a full cycle before being taken out of service. However, the vehicle is delayed further in its return trip, causing a late exit. The decision not to short-turn in the last cycle is final because short-turning has been restricted in order to prevent the strategy from forcing passengers to alight before completing their vehicle trip. The vehicle is held more than the required minimum time at the opposite terminal, for the sake of regulating headways, again assuming no further delays in the last trip. (The plan optimizer would specify less holding time if it was aware of future delays. Also, if u' had been set earlier than u'' and there were a lateness cost involved, the planner would trade off lateness and passenger costs.) The (unexpected) delay incurred in the last trip leads to a late exit.

Headway regulation appears to make late exits more likely under the schedule-free paradigm. In schedule-based operations, the 15 minutes of grace at the end of each vehicle's run cannot be used earlier. Therefore, when a vehicle is delayed, it holds only the required minimum, even when holding would improve service. The 15 minute period at the end of the run serves as a buffer for accumulated delays. Under the schedule-free paradigm, holding is used to improve service, not to adhere to a schedule or prescribed headway. The real-time planning algorithm attempts to optimally allocate any time not used to serve trips in order to adjust the timing of trips. This can lead to improved service (e.g. lower waiting times), but if too much of the floating buffer time is used early, there may not be enough left to respond to unexpected future delays. Longer holding times at terminals and the increasing separation between the schedule-based and the schedule-free space-time trajectories over time in Figure 4-15 show this happens for at least some vehicles.

Figure 4-16 shows the schedule-based, schedule-free, and scheduled space-time trajectories for the vehicle exiting latest under schedule-free operations with severe delays. In this case, the 15 minute buffer is not enough to prevent a late exit under schedule-based operations. The vehicle also exits late when operating without a schedule, for the same reason as before. We can assume that if the planning algorithm had been aware of future delays when dispatching the vehicle for its last cycle at 3:43, it would have the vehicle short-turn to prevent a late exit.

Figures 4-17 and 4-18 show how much later trips are dispatched under schedule-free operations than under schedule-based operations, in the cases of moderate and severe delays, respectively. Dispatch times under schedule-based operations are shown on the horizontal axis, and the difference between schedule-free and schedule-based dispatch times for each vehicle trip is shown on the vertical axis. Trips are compared in sequence, e.g. the third trip under each paradigm for the same vehicle. More trips begin later under the schedule-free paradigm, and the difference increases over time, reinforcing the previous observation that the buffer time between the last scheduled stop time and the latest allowed exit time, which under schedule-based operations is fixed at the end of each vehicle's run, is used gradually under the schedule-free paradigm, which increases the risk of late exits when vehicles are later delayed unexpectedly.

Systematic differences in running time predictions between the performance model and the simulation may be responsible for a few minutes of exit lateness. For example, in the performance model passengers are represented with continuous variables and it is assumed that vehicles stop at all stops, leading to errors in the predictions of dwell times at stops, and thus trip running times. This could be part of the reason some vehicles exit a few minutes late under the schedule-free paradigm, in cases of both moderate and severe delays. Stochasticity is not a contributing factor because the simulation is deterministic. The performance model, which forecasts vehicle trajectories and costs given candidate plans, could be calibrated so that estimated running times match real running times more closely. Alternatively, constraints could be modeled tighter to compensate for systematic differences; for example, the target exit times given to the plan optimizer could be 10 minutes (rather than 15 minutes) after the latest scheduled stop visit.

The difference in performance between paradigms widens under severe delays. With short-turning allowed, the mean waiting time decreases by 2.2 min-

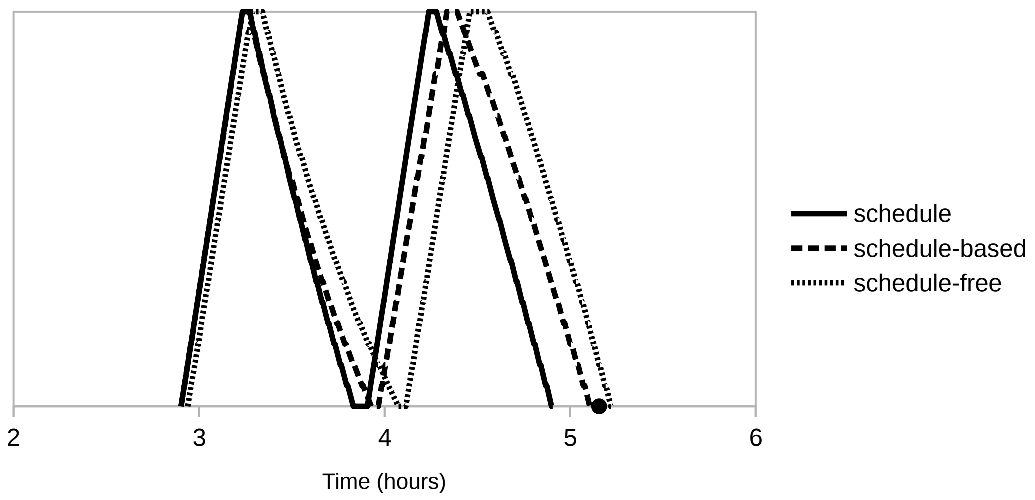


Figure 4-15: Trajectory of Vehicle with Latest Exit, Moderate Delay

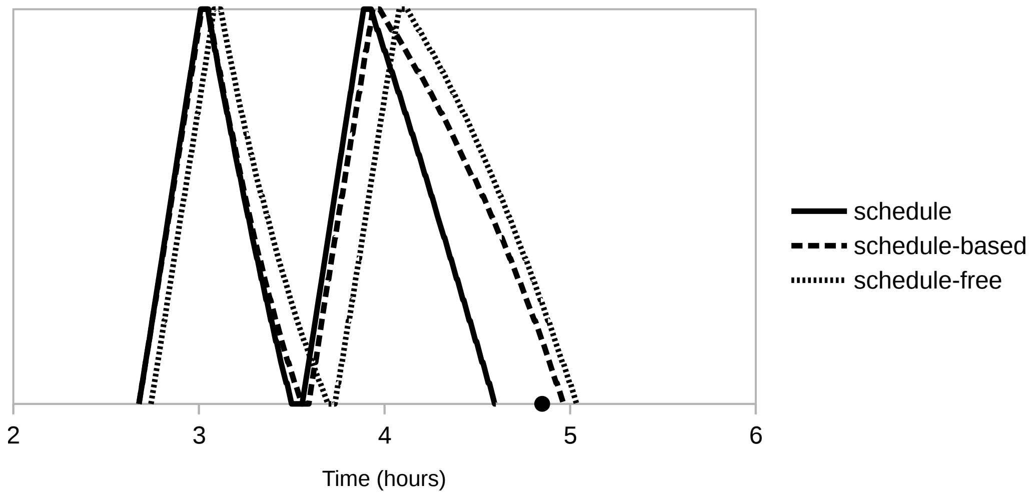


Figure 4-16: Trajectory of Vehicle with Latest Exit, Severe Delay

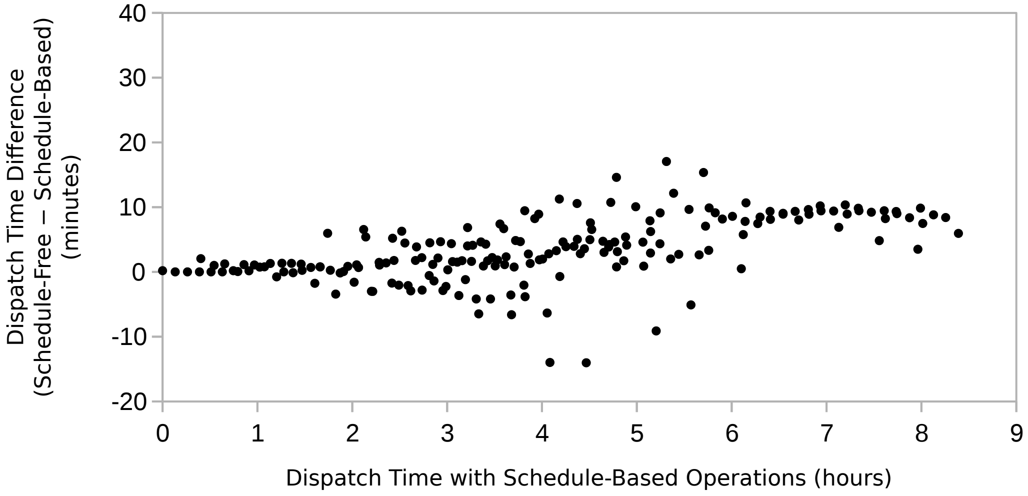


Figure 4-17: Schedule-Free Dispatch Lateness, Moderate Delay

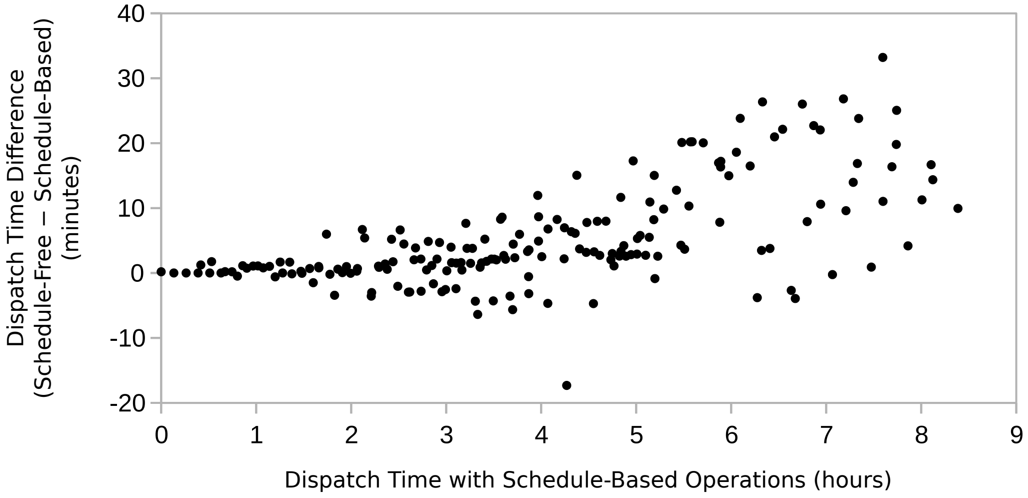


Figure 4-18: Schedule-Free Dispatch Lateness, Severe Delay

utes (35%) going from schedule-based to schedule-free operations, while mean in-vehicle times differ by only 0.1 minutes. Excess waiting time decreases by 59%. In this case, many vehicles are delayed under the schedule-based paradigm, which leads to no holding beyond the required minimum, and thus significant bunching. Vehicles are held to improve service quality under the schedule-free paradigm. The 15 minute grace period is insufficient to recover from severe delays under schedule-based operations, in spite of 16 short-turns, making 7 vehicles exit late, 6 of which exit between 4 and 10 minutes late. There are 5 more vehicles exiting late under schedule-free operations than under schedule-based operations, and the maximum exit lateness is also greater; 6 vehicles exit between 2 and 12 minutes late. Schedule-free operations lead to 4 fewer trips than scheduled, and 5 more short-turns.

It could appear paradoxical that the operations strategy serving fewer and shorter trips provides better service to passengers but leads to more late exits. The greater number of late exits happens for the reasons discussed earlier: planning unaware of future delays, combined with using buffer time much earlier to regulate headways, resulting in gradual cumulative delays with respect to the (from the real-time planning algorithm's perspective, non-existent) schedule, in addition to a smaller contribution from systematic errors in the forecasts of the performance model. Vehicles are held earlier and longer, resulting in further delays. Some vehicles encounter significant delays in their last cycle, when stop-skipping is no longer an option. Others encounter delays earlier, giving the planning algorithm the choice of either short-turning or planning fewer trips (taking the vehicle out of service earlier than originally planned). Three vehicles exit earlier than the last scheduled stop time under the schedule-free paradigm, one of which exits over 23 minutes earlier, i.e. over 38 minutes earlier than the latest allowed exit time.

In spite of serving fewer trips with more short-turning, the schedule-free paradigm achieves better headway regularity during and after the peak. Figures 4-19 and 4-20 show space-time trajectories of all vehicles between 3:00 and 8:00 for schedule-based and schedule-free operations, respectively, with short-turning allowed. Unlike the previous space-time plots, these have the first direction in the bottom half and the second (return) direction in the top half. Operating according to the schedule leads to several bunches of vehicles around 4:00, followed by many short-turns, skipping the first stops of the second direction, which has strong demand. Vehicles also short-turn under schedule-free operations, but departures from the terminal are more evenly spaced. Between 4:30 and 5:30 only 2 complete trips are offered in the peak direction under schedule-based operations, while 7 are offered under schedule-free operations.

The above tests were repeated without short-turning. Disabling short-turns simplifies the trip sequence optimization process considerably by reducing the dimensions of the trip sequence optimization problem; the only decision left to make is the number of cycles to run. Departure times must still be optimized, but the subproblem is solved far fewer times.

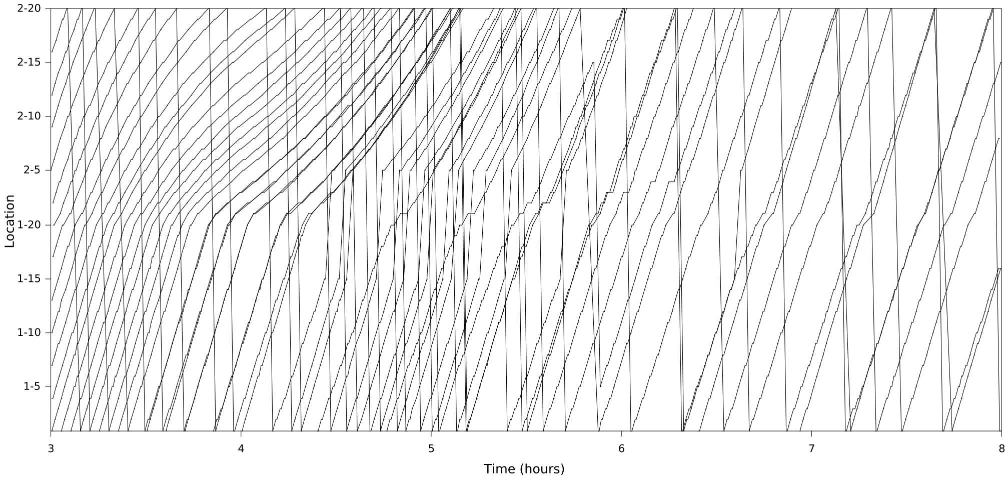


Figure 4-19: Schedule-Based Vehicle Trajectories, Severe Delay

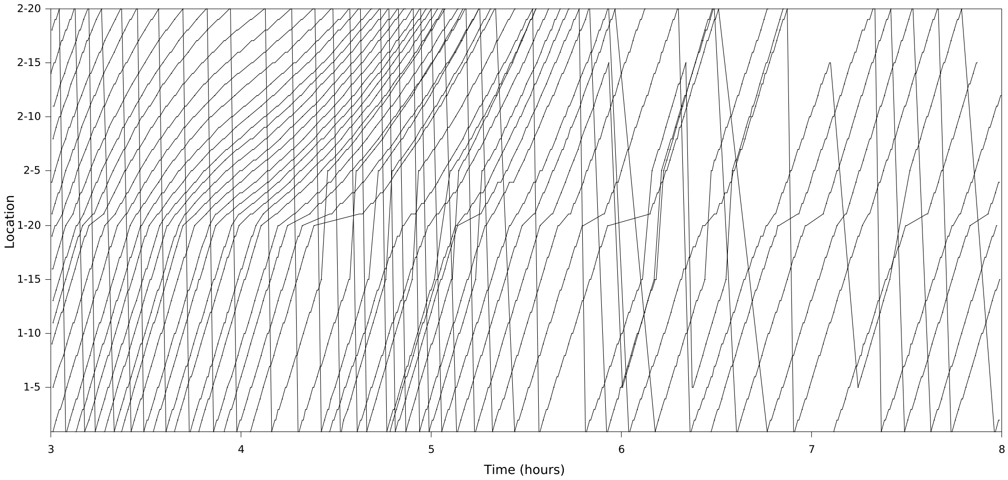


Figure 4-20: Schedule-Free Vehicle Trajectories, Severe Delay

As before, the two paradigms perform similarly without delays. In the case of the moderate delay, waiting time decreases by 0.7 minutes (20%) when operating schedule-free, and the difference in in-vehicle time is small. The waiting time savings are equivalent to an 82% drop in excess waiting time. There are 1 and 5 late exits observed with schedule-based and schedule-free operations, respectively, but the maximum exit lateness is small in both cases. Schedule-free operations result in 4 fewer trips served.

In the case of severe delays, schedule-free operations lead to fewer late exits and similar maximum exit latenesses, mean passenger waiting times, mean excess waiting times, and mean in-vehicle times. The schedule-based paradigm, as defined in this application, always serves all scheduled trips, so when short-turning is not an option, it can only absorb lateness through recovery time between trips, which in the case of severe delays is insufficient. As a result, vehicles are not held beyond the required minimum times, and bunching ensues. The schedule-free paradigm can serve fewer trips to avoid exit lateness but, as before, it plans trips assuming no additional delays, which causes exit lateness. Although 28 fewer trips are served under the schedule-free paradigm, a similar passenger cost and fewer late exits are achieved, which demonstrates how effectively headway regulation benefits passengers relative to an (in effect) uncontrolled transit line or, seen from the opposite angle, how harmful bunching can be.

Comparing mean passenger waiting times under schedule-based operations with and without short-turning, it is evident that, although short-turning does reduce late exits, applying it with a myopic strategy can significantly increase waiting times. While the myopic strategy will short-turn consecutive vehicles if they are all delayed, the schedule-free optimization approach can find less detrimental combinations of trip sequences.

Comparing the performance of the schedule-free paradigm with and without short-turning, the removal of the option to short-turn results in similar performance in the case of moderate delays, and improved performance in the case of severe delays. Schedule-free operations without short-turning result in decreased waiting time and fewer late exits than those where short-turning is allowed. This is another apparently paradoxical result, because having the option of short-turning gives the real-time planning algorithm more options, suggesting that it should always perform better. This result is probably a consequence of the simplified optimization method used for the present application. Since the full planning problem has high dimensionality, in this research we adopted a simplified decomposition method, in which trip planning problems are solved sequentially for each vehicle, based on assumed (or previously optimized) trip sequences for the rest of the vehicles. This approach prunes a large portion of the search space. It can happen that, based on initial assumptions about the trip sequences of other vehicles, a particular sequence with short-turning minimizes cost in a local univariate subspace, but that a different sequence (without short-turning) could lead to lower cost over the course of the all the individual vehicle subproblems. In other words, the optimal solution of a vehicle's subproblem, given assumed trip sequences for other vehicles, may not be the optimal choice for the general (all vehicles) problem. Since the problem without short-turning has far fewer dimensions, finding a good solution is more likely. On the other hand, the problem with short-turning may have better solutions available, though more difficult to find.

This result suggests that global optimization parallel metaheuristics could be of value, in particular at the level of decomposition of the global planning problem into sequential planning problems for each vehicle, as described in Section 4.5.1. For example, the all-vehicles trip planning problem could be solved with and without short-turning, in parallel, and the best of the two results would be selected. In addition, the individual vehicle subproblems could be solved in different orders (again, in parallel) and the best of all results would be selected.

Another possible reason why the removal of the option to short-turn leads to better performance is that plans without short-turns may be better across multiple plan updates. Plans with short-turning may lead to lower predicted cost in a particular optimization event, but perhaps plans without short-turning are more robust to unexpected future delays. Future work is needed to know if this is the case.

4.7.3 Evaluation of Feasible Sequence Set Cardinality Reduction Heuristics

Section 4.5 discusses the complexity of the real-time optimization problem and proposes heuristics that can be used to make the problem tractable and practical for real-time applications. This section explores the computational cost and feasibility of real-time operations planning using various heuristics. The objective is to select an approach that enables evaluating the feasibility and potential of the schedule-free paradigm. What is presented here is merely a first look; a more exhaustive analysis should lead to improved optimization methods.

The methodology introduces heuristics at several levels. First, the problem is decomposed into subproblems for each vehicle, separate but sequentially dependent. A ranking function determines the vehicle order. Making the problem non-combinatorial is regarded as a necessary dimensionality reduction step. Second, heuristics can be used to reduce the number of candidate trip sequences to evaluate. Third, the even headway algorithm can be used to approximate the policy of the departure time optimization problem, instead of employing a general minimization algorithm. Fourth, a combination of approximation and evaluation can be used to optimize departure times efficiently.

The following analysis focuses on sequence set cardinality reduction heuristics that play a role in the trip sequence optimization stage. These are listed in Section 4.6.3. Schedule-free operations, in the case of moderate delays with short-turning allowed, were simulated with different sets of heuristics resulting in the computation times shown in Table 4.4. Simulations were run on a computer having an Intel Core i7-3930K processor running at 3.20GHz. Departure times were optimized with the even headway algorithm in all cases. The base case, shown in the first row, skips optimization for vehicles with a feasible basic sequence (FBS), removes feasible sequences (FS) with excessive stop-skipping, and removes parent feasible sequences. (The reader is referred to Section 4.6.3 for a discussion of these heuristics. As discussed in Section 4.5, basic sequences are those composed of only end-to-end trips, i.e. without stop-skipping, and feasible sequences are those that end at a vehicle's exit location on or before the latest allowed exit time.)

In 51% of base case optimizations (one every 5 minutes), all vehicles had feasible basic sequences, so trip sequences were selected without solving the departure time subproblem. Departure times are optimized only once when this happens. At most 6 vehicles required trip sequence optimization. A maximum of 802 sequences were evaluated with the departure time subproblem for a vehicle with an infeasible basic sequence. With a mean optimization time under 10 seconds (18.7 seconds excluding instances in which all vehicles had feasible basic sequences) and a maximum of 221.2 seconds, it is feasible to plan operations in real-time with these heuristics.

The second row of Table 4.4 shows what happens when feasible trip sequences having excessive stop-skipping are not removed. For example, if a vehicle is only 5 minutes short of exiting on time with its basic sequence, the optimizer will nevertheless consider trip sequences short-turning additional times after more than 5 minutes have been saved through short-turning. The simulation was stopped before completion, when the operations plan had been updated only 14 times. In 2 of these, all vehicles had feasible basic sequences, and the optimization completed in less than 1 second. In the other 12, one to three vehicles required evaluating sequences with the departure time subproblem, and 874–6864 sequences were evaluated. Plans optimized with initial conditions that are over 20 minutes old are likely suboptimal. This demonstrates the utility of removing feasible trip sequences with excessive stop-skipping.

The third row of Table 4.4 shows what happens when feasible sequences

Table 4.4: Effect of Cardinality Reduction Heuristics on Computation Time

Cardinality Reduction Heuristics			Computation Time (s)	
Skip Vehicles with FBS	Remove FS with Excessive Stop-Skipping	Remove Parent FS	Mean	Max
•	•	•	9.3	221.2
•	—	•	1374.7	3040.6
—	•	—	40.6	234.2

are evaluated even for vehicles having feasible basic sequences. Vehicles with feasible basic sequences have no exit lateness, so the removal of sequences with excessive stop-skipping only retains sequences without short-turning, i.e. any stop-skipping is regarded as excessive in this case. Although removing parent sequences (i.e. not taking vehicles out of service when they could serve additional trips) generally makes sense, it was disabled for this case; otherwise the basic sequence is the only sequence left, and the case is equivalent to the base case. The mean computation time is 4.4 times greater than in the base case, although the maximum computation time is not much greater.

Although a case in which departure times are optimized with a general algorithm after even headway policy approximation was not tested, results from Chapter 2 can be used to estimate computation times because the holding control problem is analogous to the departure time planning problem. Departure time approximation took 0.14 seconds per sequence in the base case. Assuming that refining the approximate solution takes 1 second, it would take an estimated 1.14 seconds per sequence evaluated to optimize departure times, which is 8.2 times longer. Assuming that 96% of the computation time is spent optimizing departure times to evaluate sequences, as in the base case, updating plans would take an average of 73.4 seconds, and a maximum of 29 minutes.

Since candidate trip sequences of a single vehicle are mutually independent, they can be evaluated in parallel. In an optimistic case where there are as many processor cores as feasible trip sequences, it would be feasible to disable elimination heuristics and fine-tune departure times with a general purpose optimizer, and still update plans quickly enough. Nevertheless, all evaluations of the schedule-free paradigm presented in Section 4.7.2 use the heuristics enabled for the base case.

4.8 Concluding Remarks

High-frequency transit systems face stochastic running times and demand. Operating conditions are affected by external factors such as traffic and weather, that cannot be predicted far in advance. Operations planning typically involves scheduling, which produces a rigid plan that can be suboptimal when conditions differ from those assumed to build the schedule. This research develops a schedule-free operations planning paradigm in which operations are driven by real-time optimization. Under the new paradigm, transit systems adapt to current and expected future conditions to maintain service quality while satisfying resource constraints. The only part of operations planning that takes place before service delivery is entry and exit planning, which defines when vehicles and drivers enter and exit the system. Real-time plans are updated at short intervals (e.g. every 5 minutes). Stop-skipping strategies such as short-turning can be employed to increase fleet and driver utilization and manage overcrowding.

Adopting the schedule-free paradigm has implications on the way transit lines are operated. The new paradigm requires real-time communications between vehicles and a central computer to collect data on vehicle locations and passenger activity, and to send updated plans to vehicles. The computer used for planning models current and future states in the course of optimizing plans, keeping an estimate of waiting times, trip times, loads, headways, and other performance measures that can be shared with managers, controllers, and passengers. Routine disruptions could be handled seamlessly. Passengers do not need to change their approach to using the transit service because they do not plan to take specific scheduled trips. Drivers may experience more uncertainty in the sequence of trips run each day. Their exit times could become more certain when facing expected delays, but more uncertain when facing unexpected delays. Beyond conventional transit systems, the schedule-free paradigm could be applied to centrally plan operations in an informal setting where vehicles are independently owned and vehicle availability is only known in the short-term, as well as to transit services with autonomous fleets, which would otherwise be unduly constrained by scheduled operations.

Plan optimization is driven by a cost minimization approach capturing passenger waiting times and in-vehicle times, driver exit lateness, and solution complexity, which can be used to prevent overly complex plans that give only a marginal improvement in performance. Since the cost function is nonlinear and non-differentiable, it difficult to find globally optimal solutions. The operations planning problem is combinatorially complex, making it particularly challenging to solve in real-time. In the interest of tractability, the problem is decomposed into sequential planning problems, one per vehicle, which are solved reflecting plans for other vehicles. This subproblem is further decomposed into a trip sequence problem and a departure time subproblem. Feasible trip sequences are evaluated by solving the departure time subproblem, and the minimum cost sequence is selected. Heuristics can be used to eliminate sequences that are unlikely to be optimal, reducing the cardinality of the set of feasible sequences requiring evaluation. A discrete event performance model is used to evaluate candidate plans. A constrained even headway algorithm is used to obtain an initial point for the departure time subproblem, or it can be used without further optimization in a policy approximation approach. Dynamically modeled running times and demand are used, along with real-time vehicle locations, to estimate the current system state, and the operations plan, specifying target departure times and trip sequences which may involve stop-skipping, is updated at regular intervals.

The schedule-free paradigm is applied to a simulated transit line, in a case without delays as well as with moderate and severe delays, both with and without short-turning. Performance outcomes are compared with the schedule-based paradigm. While the two paradigms result in similar performance in the absence of delays, the schedule-free paradigm generally leads to lower passenger waiting times, but more late exits. Differences in passenger cost, number of vehicles late, magnitude of exit lateness, and amount of short-turning increase with delay magnitude. The observation of short-turning in the case without delays suggests that short-turning is sometimes planned to decrease passenger cost, perhaps by increasing frequency on a busy portion of the transit route.

Vehicles can exit late when they incur unexpected delays in their last cycle. Three factors combined lead to late exits: the unawareness of future delays when updating plans, the tendency of the real-time planning algorithm to distribute slack time throughout a vehicle's run in order to regulate headways, and the restriction on short-turning (allowing short-turning decisions to be made only between trips). The passenger cost minimization objective encourages more holding than what would be applied under scheduled operations when the line experiences delays. Holding can decrease waiting times, on the one hand, but increase the risk of exit lateness, if there are further delays, on the other. The current methodology captures the former but not the latter. Unexpected delays are probably worse than anticipated ones because they can lead to situations that are either difficult to recover from or costly to passengers. However, further research is needed to understand the benefit of knowing or predicting delay, and the potential drawbacks of complicating the strategy and making it susceptible to erroneous information. Different results can be expected if conditions changed. If the running times given to the planning algorithm reflected future delays, stop-skipping would be employed anticipating delays. Aggressive stop-skipping strategies could help prevent late exits due to unexpected delays, although passengers would at times be forced to alight before reaching their intended destination stop. This inconvenience could be modeled as an additional cost to passengers.

When short-turning is disabled, the optimizer can only prevent exit lateness by planning fewer round trips. In spite of limiting the options available to the real-time planning algorithm, the resulting plans perform similarly to those formed with short-turning available in the cases of no delay and moderate delays, while outperforming them in the case of severe delays. This is an indication that the methodology could be improved to find better combinations of trip sequences and departure times. Global optimization parallel metaheuristics, in particular at the level of decomposition of the global planning problem into sequential planning problems for each vehicle, could improve the performance of schedule-free operations further. Multiple instances of the same planning problem could be solved, in parallel, using different variations of the methodology, to then select the best outcome.

The approximation of the departure time optimization policy by the constrained even headway algorithm could be driving longer headways and waiting times, and greater number of short-turns. Combined with running time delays, high demand in the second direction makes the line reach capacity and causes passengers to be left behind at stops. Chapter 2 shows that the even headway policy is suboptimal in this case. Instead of holding to regulate headways, the optimal policy might be to maintain line capacity by running vehicles as frequently as possible while crowding persists. Headway regulation delays vehicles further, potentially making trip sequences infeasible, which triggers short-turning or planning fewer round trips in order to avoid exit lateness, which in turn further decreases capacity. Thus, higher performance could be achieved with actual optimization of departure times.

Given the complexity of the problem, a full stochastic optimization is not yet within reach. However, simple approaches can help make the planning strategy robust to uncertainty about future delays. For example, the target and maximum allowed exit times given to the planning algorithm could be changed over time. Earlier times could be given at the beginning of the day, to start with tighter constraints, and slack could be added by gradually delaying exit constraints. Alternatively, lateness cost could be specific to each vehicle, starting high to discourage early use of too much slack, and decreasing over time. The motivation behind such strategies is making plans robust by reserving some buffer time for unexpected delays, decreasing the need to revisit plans and have only bad (feasible) trip sequences to choose from. A more direct alternative is making exit time constraints a function of running times. In this case, more exit time would be made available when exogenous factors, such as traffic in a shared right of way, slow vehicles. This might reflect an operator's adaptable tolerance for lateness.

Besides proposing the schedule-free paradigm and developing its framework, this research takes what should be regarded as a first step in developing optimization methods for real-time operations planning. Results of the simple application demonstrate the feasibility and potential of schedule-free operations for high-frequency transit, but further methodological refinement and evaluation are required to ascertain the performance benefits of schedule-free operations for high-frequency transit. Future work should develop the methodology to optimize entry and exit plans, perhaps based on simulated schedule-free service under different operating conditions. It is worth exploring the modeling of driver constraints in more detail, e.g. constraints on the minimum duration of breaks between spells of work of a single driver, which introduces dependency between what is modeled as separate vehicles in this research. The potential value of strategies such as deadheading, expressing, unrestricted short-turning, and injection of spare vehicles should be investigated. The schedule-free paradigm should be evaluated in a wide range of transit services and cases in order to better understand its robustness. Applying it to a real service could lead to a better understanding of its implications on human factors.

Chapter 5

Conclusion

This research advances the state-of-the-art in operations planning and control for high-frequency transit. The widespread use of automated data collection systems and communication technologies enables real-time capture of events recorded by vehicles, fare gates, etc. These streams of data can be centrally processed and combined with historical data to model current and future performance. The information obtained from data fusion and modeling can enhance the effectiveness of control strategies and planning.

This research focuses on enhancements brought by modeling future states with dynamic running times and demand, which capture both typical daily patterns of exogenous changes in operations and changes due to particular events, such as traffic accidents or demand surges characteristic of large social gatherings such as concerts and sport contests. Since it is not possible to utilize real-time information to shape operations plans prepared before service delivery, a schedule-free paradigm for operations planning is developed, in which trip and stop level vehicle activities are planned in real-time, taking advantage of the latest available information. The feasibility and effectiveness of operating high-frequency transit with the proposed models is demonstrated through applications to simulated transit services.

5.1 Summary

Operations control is an important means of improving service quality in high-frequency public transport systems. It is based on continuous monitoring of the system and supply-side interventions with the aim of improving service quality. Holding, the most commonly employed intervention, consists of intentionally delaying a vehicle, possibly at the expense of extending trip times for passengers on board, in order to reduce the waiting time of passengers who will board downstream. Past research has evaluated the effectiveness of different control strategies, including ones not requiring real-time information, others based on local information and myopic heuristics, and yet others based on optimization models involving future system state predictions, assuming static running times and demand.

Chapter 2 presents an optimization-based holding control model that captures dynamic running times and demand. This enables control that reacts not only to what is already known from observing the current state, but also to what can be anticipated based on historical experience and knowledge of future conditions. For instance, information about a demand increase expected 20 minutes into the future due to rush hour operations can affect which strategies are optimal. Awareness of upcoming maintenance work enables anticipating unusually long running times in part of a line, creating the opportunity to preemptively control in a manner that minimizes delays to passengers. This form of anticipatory control can lead to improved performance with respect to strategies that react only after disruptions materialize.

Optimization is based on a deterministic rolling-horizon performance model that takes as inputs general (e.g. nonlinear or piecewise) running time and demand functions, in addition to vehicle locations, load estimates, and the number of passengers currently waiting at stops. The model estimates future system states, including departure times and loads, from which passenger waiting and in-vehicle times are derived. Passenger demand is modeled continuously, and it is assumed that vehicles stop at every stop and do not overtake. The model can nevertheless be applied to control stochastic systems in which vehicles do not stop at every stop and can overtake. The optimization model minimizes mean passenger cost, which combines waiting time and in-vehicle time, normalized by the total number of boarding passengers, subject to constraints on maximum allowable holding times at each stop as well as constraints describing vehicle movement and passenger activity. If necessary to obtain solutions quickly, the dimensionality of the problem can be reduced through a lower number of control points and exclusive optimization of holding times for a control subset of vehicles.

The new control strategy is evaluated by comparing the performance of a simulated high-frequency bus service controlled with four different holding strategies: target headway, even headway, optimization-based with static inputs, and the new optimization-based with dynamic inputs. Six cases are considered, including ones with static and dynamic running times and demand, with low and high crowding. Performance is measured in terms of mean passenger cost, mean headway and headway variability, and crowding at stops. Optimization-based control strategies lead to similar or better performance than the target headway and even headway strategies. The dynamic optimization strategy outperforms the static optimization strategy in cases where, due to running time or demand dynamics, the line becomes overcrowded and passengers are left behind by (full) vehicles. In the cases involving dynamic demand and high crowding, decreases of 6.4% and 7.2% in mean passenger cost (twice the waiting time plus in-vehicle time) and 15.8% and 25.0% in excess waiting time were observed with respect to the optimization strategy using static inputs. In cases having dynamic running times but not significant overcrowding, the dynamic strategy modestly outperforms its static equivalent.

Headway regulation is the principal mechanism by which the new strategy improves performance. The optimal control policy of an overcrowded transit line may be different from minimizing headway regularity. A large part of the performance improvement is due to more accurate estimates of the current state, while the remainder comes from estimating future states with dynamic running times and demand. Computation times of the new strategies are suitable for real-time application.

Chapter 3 builds upon the model developed in Chapter 2 by presenting a framework for real-time holding control with information about events. Events such as road congestion caused by traffic accidents, rail signal failures, and medical emergencies in vehicles or stations are unpredictable. Other events such as concerts and sport contests cause predictable surges in demand, resulting in short-term local congestion. Regardless, information about events can be considered in the generation of a control response. Although events such as traffic accidents cannot be predicted, the detection of the event along with an estimate of its duration based on experience and real-time updates on the progress towards resolution can be used to predict congestion and how it might affect transit service. The end of a concert or sports event is followed by an increase in traffic and demand for transit service. The time at which the event ends, associated traffic delays, and number of people from the event who will take transit can be estimated with information about event attendance, real-time updates on the progress of the event, and observations of operations during similar events in the past. These predictions can be used to apply control preemptively instead of waiting until service is disrupted enough that a problem can be detected. Since events generate transients in running times and demand, the control model developed in Chapter 2 is a key enabler.

Controlling transit operations capturing event-driven dynamics can potentially improve the effectiveness of control, reducing waiting times and trip times for passengers during an event. The performance benefit of a strategy that captures event-driven dynamics, i.e. an informed strategy, is derived from its information advantage over a strategy that neglects event-driven dynamics, i.e. a naive strategy. Having more realistic predictions of running times and demand can lead to better predictions of future system states (and their cost to passengers) under different control scenarios, ultimately leading to more effective control policies. In order to control a transit line considering the effects of events, it is necessary to become aware of events, gather relevant data, model future operating conditions, and optimize holding times capturing event-driven dynamics.

The naive and informed strategies, both based on the optimization model of Chapter 2, are compared through simulated transit service. Two cases are considered: an unforeseen event (e.g. a traffic accident) causing a link between stops to be blocked for a short period during rush hour, and a foreseen demand surge (e.g. due to a concert ending). Since the advantage of the informed model grows as information about the event is known farther in advance, the performance improves more with foreseen events than with unforeseen events.

Preemptive holding suggested by the informed model increases passenger cost shortly before a foreseen event affects service, but decreases passenger cost thereafter, for an overall net benefit. In the specific case presented, an 18% cost reduction was observed for passengers arriving at the two surge stops during the demand surge, and the peak number of passengers waiting at the affected stops decreased by 22%. While the results are fairly robust to errors in the estimated surge magnitude, controlling with errors in event timing can be counterproductive. Greater benefits might be possible with stop-skipping strategies such as short-turning, but this was not investigated.

One characteristic of advanced real-time control strategies, including the ones presented in Chapters 2 and 3, is that they function independently of the operations plan. On the one hand, this provides them with the flexibility required to regulate service. On the other, it can put control policies that benefit passengers at odds with operator constraints, such as those related to driver exit times. Operations planning for high-frequency transit remains heavily focused on schedules, which are deterministic and constrain the availability of vehicles and crew.

Chapter 4 proposes and develops the framework for a schedule-free paradigm for high-frequency transit, in which operations planning is driven by real-time optimization, allowing transit systems to adapt to current and expected future conditions to maintain service quality while satisfying resource constraints. An operator following the new paradigm would allocate vehicles and drivers to a service (or set of services) by time of day. Only entry and exit times and locations would be specified a-priori. Trip and stop level planning would take place while service is being delivered, reflecting current and expected future conditions, as well as changes in the available number of vehicles and drivers. The plan optimization model can be given flexibility by allowing it to plan short trips, with strategies such as short-turning and deadheading. Short trips can be offered when there is not enough time remaining in a driver's shift for a full trip, and to improve service by increasing frequency in a targeted manner.

Schedule-free operations planning relies on automated data collection systems and communications and information technologies to collect and process data, to optimize plans, and to send updated plans to vehicles. A combination of historical and real-time data is used to model the current and future system states under potential plans. Since schedules are deeply entrenched in transit organizations, the schedule-free paradigm has implications on methods and practices followed for incident management, performance measurement, and passenger information provision. Although passengers would not have to change their approach to taking transit (or even be aware of the new paradigm) the experience of drivers and managers can change significantly when plans are updated continuously. The schedule-free paradigm could be useful for operations with an uncertain number of vehicles and drivers, as well as for operations with autonomous (driverless) fleets.

The real-time planning problem formulated in Chapter 4 seeks to minimize a function of passenger cost, driver exit lateness cost, and plan complexity cost, subject to constraints on the usage of vehicles and drivers. Complexity cost can be added as a disincentive for plans involving stop-skipping in exchange for only marginal passenger and exit lateness benefits. The lateness policy can be modeled combining hard constraints and lateness costs. For example, there might be a target exit time through which a vehicle and driver serve without cost, and a maximum allowable exit time, which prevents plans with excessive exit lateness. A nonlinear cost function can be used to model lateness cost between these two parameters.

The full planning problem has combinatorial complexity and is intractable for problems of realistic size. Particularly for real-time application, a fast solution is required. The simplified approach adopted in this dissertation is to decompose the general problem into sequential subproblems for each vehicle. This separation does not imply independence, as the trip sequences planned for the entire active fleet are taken into account when optimizing the trip sequences of individual vehicles. Initially assumed trip sequences do not allow any stop-skipping, with the exception of previously planned stop-skipping required to complete the current trip. A ranking function is used to guide the order in which vehicle plans are optimized. Each vehicle's optimization problem is further decomposed into a trip sequence problem and a departure time problem analogous to the holding control problem discussed in Chapters 2 and 3. It is difficult to establish mathematical relationships between feasible sequences, so they are enumerated and evaluated by optimizing departure times for each. Heuristics can be used to reduce the number of feasible sequences to evaluate. For instance, trip sequences with excessive short-turning or needlessly early exits can be removed. A constrained even headway algorithm is used to optimize departure times. The problem itself and the approach followed make it difficult to find globally optimal solutions.

An application of the schedule-free paradigm to a simulated transit line is used to evaluate its feasibility and potential. The performance of the transit line under the schedule-free paradigm is compared to that under the schedule-based paradigm, in a case without delays as well as with moderate and severe delays. Short-turning is allowed in both directions, but there are no deadheading or expressing options. Running times without delays are assumed both by the schedule in schedule-based operations and the real-time planning model in schedule-free operations. Hence, delays are disruptive under both paradigms. The two paradigms result in similar performance in the absence of delays. Except in the case of severe delays without short-turning, the schedule-free paradigm reduces mean passenger waiting times but increases vehicle exit lateness, particularly for vehicles delayed in their last round trip. Due to restrictions placed on short-turning in order to prevent forced alightings, in addition to increased holding for headway regulation in response to delays, vehicles become more vulnerable to unexpected delays, which can be difficult to recover from (e.g. when a vehicle is delayed in its last trip and the original plan was already tight with respect to satisfying the exit time constraint) or harmful to passengers (e.g. when a vehicle must be short-turned or taken out of service early because there is no other way to satisfy the exit constraint). However, headway regulation improves service quality, even when it results in fewer trips or a greater number of short-turns. Since there is a trade-off between improving service quality for passengers and satisfying exit constraints, it is not possible with the present results to establish which of the two paradigms is generally better.

Removing the option of short-turning significantly restricts the flexibility of the schedule-free planning model, particularly because the dimensionality of the trip sequence problem reduces drastically. This has the benefit of making optimization much more tractable and easier to find good solutions, but has the drawback of removing potentially better plans. For example, short-turning could help prevent a late exit without the more drastic alternative of taking a vehicle out of service much earlier than its latest allowed exit time. It could also help manage crowding by increasing capacity locally. Results show the new paradigm performing better without the short-turning option, but this could be due to the limitations of the simplified methodology.

5.2 Contributions

This research has advanced, through new frameworks and methods, the utilization of historical and real-time information to improve high-frequency public transit operations planning and control. The major contributions of this research are:

1. The formulation of a holding control model capturing dynamic running times and demand, allowing holding policies that reflect anticipated changes in running times and demand.

2. Knowledge that information about dynamics, when applied to the holding control problem, can significantly improve the performance of highly crowded high-frequency transit services, particularly when crowding is more accurately modeled using information about dynamics, but that it does not result in significant performance improvements in less crowded systems.

3. A framework for controlling a high-frequency transit service with information about foreseen and unforeseen events.

4. Evidence that information about events, particularly anticipated events, can lead to significant performance benefits, but that erroneous information, particularly about event timing, can significantly diminish this benefit, potentially to the point of counterproductivity.

5. A schedule-free paradigm for high-frequency transit operations planning, in which most operations planning decisions are made in real-time with the latest available information.

6. A discussion of potential methodological approaches to real-time schedule-free planning, and a specific simplified methodology with holding and stop-skipping strategies, which lays the groundwork for future improvements.

7. Evidence from simulation experiments that the schedule-free paradigm is feasible, and that, when the real-time planning model is unaware of delays, it generally leads to lower mean passenger waiting times but a greater number of late exits. Results suggest that the methodology could be improved.

5.3 Future Work

In exploring holding control with dynamics, holding control with information about events, and developing a schedule-free paradigm for high-frequency transit operations planning, many new questions and opportunities for future research have arisen.

Stochasticity While this research has shed light on the value of information about dynamics to improve performance, the potential value of modeling stochasticity has yet to be explored. Recognizing stochasticity of running times and demand, both within-day and across days, as well as uncertainty in the timing, duration, and magnitude of events, should make control strategies and real-time plans more robust. While it may be difficult to model both dynamics and stochasticity in rich detail, simple approaches could be used to bring elements of stochasticity into the optimization models for holding and schedule-free operations planning. For example, inputs on lower and upper bounds of running times and passenger arrival rates could be used to forecast worst-case and best-case scenarios in addition to expected future states. Under schedule-free operations planning, the exit time constraints could be modeled as functions of uncertainty or its proxies, such as time remaining in service. For example, exit times might be modeled more tightly at the beginning of the day in order to buffer potential unexpected delays later in the day.

Information Modeling Once research shows the value of some types of information, it becomes worthwhile to investigate methods for information modeling, i.e. ways of generating information from data. Information modeling involves combining data from multiple uncertain sources, analyzing the data to learn about a transit system, and applying that knowledge to improve performance. Analysis should focus not only on understanding system states at different points in time, but also on identifying patterns and trends, such as differences due to days of the week, weather, and seasons. The role of errors and delays in receiving data and of modeling errors that produce false information from good data is not well understood, particularly in the realm of transit operations.

Control Strategies Performance improvements achievable with better information are limited by the types of strategies available. This research focused on holding for control, and holding combined with a benign form of short-turning for schedule-free operations planning. Allowing short-turning decisions only between trips means that passengers are never forced to alight before reaching their destination. The methodology allows modeling deadheading and benign forms of expressing and other stop-skipping strategies, but these were not tested. Unrestricted short-turning and expressing might unleash greater performance improvements, but they would require modeling their inconvenience to passengers. In the context of schedule-free operations, real-time interlining and limited stop service decisions could be explored. Future research could shed light on additional benefits brought by multi-line or network level optimization, as well as combining supply side interventions with real-time demand management. The latter would involve models not only for estimating current and future states given some demand, but also real-time demand models that predict how transit riders would react to information provided to them, including service alerts and personalized suggestions.

The schedule-free paradigm introduced in this research provides several avenues for further research. Much more methodological work is needed, followed by extensive experimentation to assess the true potential of the new paradigm under a wide range of operating and demand conditions.

Entry and Exit Plan Optimization Entry and exit times of vehicles and drivers, decided before service delivery, are an important part of schedule-free operation, but this research focused only on the real-time planning part. Entry and exit times can be obtained from a schedule when a transit service transitions from schedule-based to schedule-free operations, as assumed in the application presented in Chapter 4, but this is unlikely to be optimal. Performance should improve if entry and exit times are optimized based on simulated schedule-free operations under varying conditions, i.e. capturing stochasticity within and across days. This optimization might be driven by a cost function reflecting minimum performance requirements as well as passenger and operator costs. Minimum performance standards could include maximum allowed waiting times and loads by time of day. Passenger costs could reflect waiting and in-vehicle time, as well as reliability. Operator costs would capture number of drivers and vehicles, effort (e.g. measured in driver hours, vehicle-hours, and vehicle-miles), and include factors such as labor agreements and overtime compensation requirements.

Real-Time Planning Methodology More work is also needed to improve the optimization methodology used for real-time operations planning. Section 4.5 considers potential approaches not yet implemented or tested. Results suggest that departure time optimization (rather than even headway policy approximation) might improve the performance of schedule-free operations by not holding to regulate headways when the line is (or is predicted to be) overcrowded. The full problem is intractable, so it must be simplified and its dimensionality reduced for real-time applications, but there is opportunity to test more sophisticated approaches, such as metaheuristics for solving several variations of each problem in parallel and selecting the best of all optimization outcomes. This would allow trying different initial assumptions, vehicle optimization orders, and allowed stop-skipping strategies. Several passes through all vehicles (rather than a single pass) could improve the optimization outcome by increasing the probability of finding a locally optimal plan.

Driver Constraints Driver constraints could be modeled in greater detail. For simplicity, this research lumps vehicles and drivers as units of resource. For example, two pieces of work by the same driver would be modeled as two vehicles with planned entry and exit times. This approach does not capture the relationship between pieces of work by the same driver. For example, there may be constraints on the minimum duration of breaks, maximum length of time on duty, etc.. Some of this can be modeled by updating, in real time, the entry and exit times of second pieces of work. For example, a driver's second piece of work might be delayed by the exit lateness from his first piece of work. Another potential enhancement is modeling multiple drivers using a single vehicle (at different times) and the utilization of spare drivers to prevent exit lateness without relying on stop-skipping. This might require real-time assignment of drivers to vehicles, which would complicate the optimization.

Autonomous Fleets The application of the schedule-free paradigm to transit services served by autonomous driverless fleets is another potentially fruitful area of research. Scheduled operations inadequately constrain the potential flexibility of autonomous vehicles. Since exit constraints are not as important with autonomous fleets, real-time optimization can focus on passenger service quality. In addition, vehicle entries and exits could be determined in real-time rather than before service delivery, an opportunity to base real-time decisions directly on the service plan and minimum performance requirements. Vehicles could be brought into and out of service in response to overcrowding and other events.

Organizing Informal Systems It would also be interesting to study the potential application of the schedule-free paradigm to informal systems in which vehicles are owned by individual drivers. In such systems, the availability of vehicles is uncertain. Rather than preparing an a-priori plan of vehicle entries and exits, drivers could use Internet-enabled mobile devices to announce their availability (and commitment to operate for some period) on short notice. The schedule-free paradigm could be applied to centrally plan operations with these freelance drivers, potentially improving performance with respect to a system where drivers choose what trips to serve using myopic strategies, perhaps aiming to maximize fare revenue. Research might shed light on ways of aligning personal and societal goals based on contract structure and economic incentives.

Different Operating Environments The schedule-free paradigm should be tested on transit systems with different running time and demand assumptions, to learn how its behavior and achieved performance change in response to these characteristics. The transit system used for the application in Chapter 4 has a high rate of passenger arrivals at the first stops in each direction, but short-turning may be more favorable in systems having stronger demand for travel within the central portion of the route rather than to or from the ends. Some systems have a few stops with much higher demand than the rest, e.g. at interchange points.

Real Transit Services New lessons can probably be learned from applying the new control strategies and operations planning paradigm to real transit services. Although simulation modeling has the benefit of evaluating new strategies and concepts in a controlled environment, which is advantageous to scientifically quantify performance benefits, complex elements can be left out, such as correlations in running time across drivers and segments of a route. Operating real transit service under the schedule-free paradigm would require calibrating the performance and optimization models used to generate plans.

Pathway to Schedule-Free Operations More research is needed to improve the methodological aspects of the schedule-free paradigm before it can be put into practice, but in parallel with that effort it is also necessary to study other barriers, such as institutional resistance, contractual structure, and human factors. For example, what incentives and penalties should be included in a contract for high-frequency bus service operated without schedules by a private operator on behalf of a public agency in charge of service planning and contract management? How should performance be measured? Concepts from behavioral economics could be applied to study human factors. For example, what is the difference between a driver following a fixed schedule and one following frequently updated real-time instructions? Does a fixed schedule offer some comfort to drivers, supervisors, and planners? How do different communications technologies affect driver compliance? Questions of this nature should be asked for all stakeholders in transit operations, including drivers, dispatchers, controllers, supervisors, planners, and contract designers.

Bibliography

Abkowitz, M. and M. Lepofsky (1990). Implementing Headway-Based Reliability Control on Transit Routes. *Journal of Transportation Engineering 116*(1), 49–63.

Adamski, A. and A. Turnau (1998). Simulation support tool for real-time dispatching control in public transport. *Transportation Research Part A: Policy and Practice 32*(2), 73–87.

Adenso-Díaz, B., M. O. González, and P. González-Torre (1999). On-line timetable re-scheduling in regional train services. *Transportation Research Part B: Methodological 33*(6), 387–398.

Barnett, A. (1974). On controlling randomness in transit operations. *Transportation Science 8*(2), 102–116.

Bartholdi, J. J. and D. D. Eisenstein (2012). A self-coördinating bus route to resist bus bunching. *Transportation Research 46B*(4), 481–491.

Boyle, D. K. (2009). *TCRP Report 135: Controlling System Costs: Basic and Advanced Scheduling Manuals and Contemporary Issues in Transit Scheduling.*

Cats, O., A. N. Larijani, H. N. Koutsopoulos, and W. Burghout (2011). Impacts of Holding Control Strategies on Transit Performance: Bus Simulation Model Analysis. *Transportation Research Record* (2216), 51–58.

Ceder, A. (2007). *Public transit planning and operation: theory, modeling and practice.* Elsevier, Butterworth-Heinemann.

Chandrasekar, P., R. L. Cheu, and H. C. Chin (2002). Simulation Evaluation of Route-Based Control of Bus Operations. *Journal of Transportation Engineering 128*(6).

Chen, Q., E. Adida, and J. Lin (2013). Implementation of an iterative headway-based bus holding strategy with real-time information. *Public Transport 4*(3), 165–186.

Corman, F., A. D'Ariano, D. Pacciarelli, and M. Pranzo (2010). A tabu search algorithm for rerouting trains during rail operations. *Transportation Research Part B: Methodological 44*(1), 175–192.

Corman, F., A. D'Ariano, D. Pacciarelli, and M. Pranzo (2012). Bi-objective conflict detection and resolution in railway traffic management. *Transportation Research Part C: Emerging Technologies 20*(1), 79–94. Special issue on Optimization in Public Transport+ISTT2011 Special issue on Optimization in Public Transport+International Symposium on Transportation and Traffic Theory (ISTTT), Berkeley, California, July 18–20, 2011.

Cortés, C. E., S. Jara-Díaz, and A. Tirachini (2011). Integrating short turning and deadheading in the optimization of transit services. *Transportation Research Part A: Policy and Practice 45*(5), 419–434.

Şahin, . (1999). Railway traffic control and train scheduling based onintertrain conflict management. *Transportation Research Part B: Methodological 33*(7), 511–534.

Daganzo, C. F. (2009). A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. *Transportation Research Part B: Methodological 43*(10), 913–921.

Daganzo, C. F. and J. Pilachowski (2011). Reducing Bunching with Bus-to-Bus Cooperation. *Transportation Research 45B*(1), 267–277.

D'Ariano, A., D. Pacciarelli, and M. Pranzo (2007). A branch and bound algorithm for scheduling trains in a railway network. *European Journal of Operational Research 183*(2), 643–657.

D'Ariano, A., D. Pacciarelli, and M. Pranzo (2008). Assessment of flexible timetables in real-time traffic management of a railway bottleneck. *Transportation Research Part C: Emerging Technologies 16*(2), 232–245.

Delgado, F., J. C. Muñoz, and R. Giesen (2012). How much can holding and/or limiting boarding improve transit performance? *Transportation Research Part B: Methodological 46*(9), 1202–1217.

Delgado, F., J. C. Muñoz, R. Giesen, and A. Cipriano (2009). Real-Time Control of Buses in a Transit Corridor Based on Vehicle Holding and Boarding Limits. *Transportation Research Record 2090*(1), 59–67.

Desaulniers, G. and M. Hickman (2007). Public Transit. Handbooks in OR & MS 14: Transportation (C. Barnhart, G. Laporte, eds.) 69–127.

Dessouky, M., R. Hall, L. Zhang, and A. Singh (2003). Real-time control of buses for schedule coordination at a terminal. *Transportation Research Part A: Policy and Practice 37*(2), 145–164.

Eberlein, X. J. (1995). *Real-Time Control Strategies in Transit Operations: Models and Analysis*. Ph. D. thesis, Massachusetts Institute of Technology.

Eberlein, X. J., N. H. Wilson, and D. Bernstein (2001). The Holding Problem with Real–Time Information Available. *Transportation science 35*(1), 1–18.

Gordon, J. B., H. N. Koutsopoulos, N. H. Wilson, and J. P. Attanucci (2013). Automated Inference of Linked Transit Journeys in London Using Fare-Transaction and Vehicle-Location Data. In *Transportation Research Board 92nd Annual Meeting*, Number 13-0740.

Graham, D. J. (2007a). Agglomeration, productivity and transport investment. *Journal of transport economics and policy (JTEP) 41*(3), 317–343.

Graham, D. J. (2007b). Variable returns to agglomeration and the effect of road traffic congestion. *Journal of Urban Economics 62*(1), 103–120.

Huisman, D. (2007). A column generation approach for the rail crew rescheduling problem. *European Journal of Operational Research 180*(1), 163–173.

Huisman, D. and A. P. Wagelmans (2006). A solution approach for dynamic vehicle and crew scheduling. *European Journal of Operational Research 172*(2), 453–471.

Hymel, K. (2009). Does traffic congestion reduce employment growth? *Journal of Urban Economics 65*(2), 127–135.

Krasemann, J. T. (2012). Design of an effective algorithm for fast response to the re-scheduling of railway traffic during disturbances. *Transportation Research Part C: Emerging Technologies 20*(1), 62–78. Special issue on Optimization in Public Transport+ISTT2011 Special issue on Optimization in Public Transport+International Symposium on Transportation and Traffic Theory (ISTTT), Berkeley, California, July 18–20, 2011.

Leiva, C., J. C. Muñoz, R. Giesen, and H. Larrain (2010). Design of limited-stop services for an urban bus corridor with capacity constraints. *Transportation Research Part B: Methodological 44*(10), 1186–1201.

Mazzarello, M. and E. Ottaviani (2007). A traffic management system for real-time traffic optimisation in railways. *Transportation Research Part B: Methodological 41*(2), 246–274. Advanced Modelling of Train Operations in Stations and Networks.

Mesquita, M. and A. Paias (2008). Set partitioning/covering-based approaches for the integrated vehicle and crew scheduling problem. *Computers & Operations Research 35*(5), 1562–1575. Part Special Issue: Algorithms and Computational Methods in Feasibility and Infeasibility.

MTA New York City Transit (2014). Subway and Bus Ridership. `http://web.mta.info/nyct/facts/ridership/index.htm`. Retrieved on December 29, 2014.

Mukhopadhyay, S. C. and N. Suryadevara (2014). *Internet of Things: Challenges and Opportunities*. Springer.

Muñoz, J. C., C. E. Cortés, R. Giesen, D. Sáez, F. Delgado, F. Valencia, and A. Cipriano (2013). Comparison of dynamic control strategies for transit operations. *Transportation Research Part C: Emerging Technologies 28*, 101–113.

Neff, J. and M. Dickens (2013, October). *APTA 2013 Public Transportation Fact Book* (64 ed.). Washington, D.C.: American Public Transportation Association.

Newell, G. F. (1974). Control of Pairing of Vehicles on a Public Transportation Route, Two Vehicles, One Control Point. *Transportation Science 8*(3), 248–264.

O'Dell, S. W. and N. H. Wilson (1999). Optimal real-time control strategies for rail transit operations during disruptions. In *Computer-aided transit scheduling*, pp. 299–323. Springer.

Osorio, C. and M. Bierlaire (2013). A Simulation-Based Optimization Framework for Urban Transportation Problems. *Operations Research 61*(6), 1333–1345.

Osuna, E. E. and G. F. Newell (1972). Control Strategies for an Idealized Public Transportation System. *Transportation Science 6*(1), 52–72.

Powell, M. J. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*.

Puong, A. and N. H. Wilson (2008). A train holding model for urban rail transit systems. In *Computer-aided Systems in Public Transport*, pp. 319–337. Springer.

Rezanova, N. J. and D. M. Ryan (2010). The train driver recovery problem—A set partitioning based model and solution method. *Computers & Operations Research 37*(5), 845–856. Disruption Management.

Rodriguez, J. (2007). A constraint programming model for real-time train scheduling at junctions. *Transportation Research Part B: Methodological 41*(2), 231–245. Advanced Modelling of Train Operations in Stations and Networks.

Rossetti, M. D. and T. Turitto (1998). Comparing static and dynamic threshold based control strategies. *Transportation Research Part A: Policy and Practice 32*(8), 607–620.

Sáez, D., C. E. Cortés, F. Milla, A. Núñez, A. Tirachini, and M. Riquelme (2012). Hybrid predictive control strategy for a public transport system with uncertain demand. *Transportmetrica 8*(1), 61–86.

Sánchez-Martínez, G. E. (2012). Running Time Variability and Resource Allocation: A Data-Driven Analysis of High-Frequency Bus Operations. Master's thesis, Massachusetts Institute of Technology.

Shen, S. and N. H. M. Wilson (2001). An Optimal Integrated Real-Time Disruption Control Model for Rail Transit Systems. *Computer-Aided Scheduling of Public Transport*, 335–363.

Site, P. D. and F. Filippi (1998). Service optimization for bus corridors with short-turn strategies and variable vehicle size. *Transportation Research Part A: Policy and Practice 32*(1), 19–38.

Sun, A. and M. Hickman (2008). The holding problem at multiple holding stations. In *Computer-aided systems in public transport*, pp. 339–359. Springer.

Törnquist, J. and J. A. Persson (2007). N-tracked railway traffic re-scheduling during disturbances. *Transportation Research Part B: Methodological 41*(3), 342–362.

Transport for London (2014). Annual Report and Statement of Accounts: 2013/14. `https://www.tfl.gov.uk/cdn/static/cms/documents/annual-report-2013-14.pdf`.

Turnquist, M. A. and S. W. Blume (1980). Evaluating potential effectiveness of headway control strategies for transit systems. *Transportation Research Record* (746).

UITP (2013). Observatory of Automated Metros World Atlas Report, 2013. Technical report, UITP Automated Metros Observatory.

United Nations (2014). The World Population Situation in 2014: A Concise Report. Technical report, United Nations, Department of Economic and Social Affairs, Population Division, New York.

Valouxis, C. and E. Housos (2002). Combined bus and driver scheduling. *Computers & Operations Research 29*(3), 243–259.

Veelenturf, L. P., D. Potthoff, D. Huisman, and L. G. Kroon (2012). Railway crew rescheduling with retiming. *Transportation Research Part C: Emerging Technologies 20*(1), 95–110. Special issue on Optimization in Public Transport+ISTT2011 Special issue on Optimization in Public

Transport+International Symposium on Transportation and Traffic Theory (ISTTT), Berkeley, California, July 18–20, 2011.

Vuchic, V. R. (2005). *Urban transit: operations, planning, and economics.*

Walker, C. G., J. N. Snowdon, and D. M. Ryan (2005). Simultaneous disruption recovery of a train timetable and crew roster in real time. *Computers & Operations Research 32*(8), 2077–2094.

Xuan, Y., J. Argote, and C. F. Daganzo (2011). Dynamic Bus Holding Strategies for Schedule Reliability: Optimal Linear Control and Performance Analysis. *Transportation Research 45B*(10), 1831–1845.

Yu, B. and Z. Yang (2009). A dynamic holding strategy in public transit systems with real-time information. *Applied Intelligence 31*(1), 69–80.

Zhao, J., S. Bukkapatnam, and M. M. Dessouky (2003). Distributed architecture for real-time coordination of bus holding in transit networks. *Intelligent Transportation Systems, IEEE Transactions on 4*(1), 43–51.