CASPT 2015

Schedule-Free High-Frequency Transit Operations

Gabriel E. Sánchez-Martínez · Nigel H. M. Wilson · Haris N. Koutsopoulos

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. 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.

Keywords high-frequency transit · schedule-free · real-time control

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.

Gabriel E. Sánchez-Martínez
77 Massachusetts Ave. Room 1-235, Cambridge, MA 02139
Tel.: +1-617-258-0699
E-mail: gsanmar@mit.edu

Nigel H. M. Wilson
77 Massachusett Ave. Room 1-238, Cambridge, MA 02139
Tel.: +1-617-253-5046
E-mail: nhmw@mit.edu

Haris N. Koutsopoulos
360 Huntington Avenue, Boston, MA 02115
Tel.: +1-617-373-2444
E-mail: h.koutsopoulos@neu.edu

1 Introduction

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 paper 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.

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, but instead expect to wait a short time after their arrival at a stop (or station). 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 2 reviews the literature, Section 3 presents the framework, Section 4 formulates the real-time trip planning problem, Section 5 presents a simple initial methodology, Section 6 applies the framework and methodology to a simulated transit service, and Section 7 draws conclusions.

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 Sánchez-Martínez, 2015).

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 Sánchez-Martínez, 2015). 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 by controlling speeds and considering rerouting. 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 regenerating timetables to resolve conflicts, with the goal of minimizing delays with respect to the original timetable. Rezanova and Ryan (2010) focus on recovering the train driver schedule 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. 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 to minimize cost when changes to the timetable or vehicle schedule have made the original crew schedule infeasible, e.g. during 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.

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.

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. Figures 1 and 2 illustrate the two paradigms. Under the schedule-based paradigm the operations plan is fully defined, and therefore fixed, before service delivery. Operations planning involves timetable development and vehicle and crew scheduling. Timetables specify vehicle departure times from stops or stations, reflecting the service frequencies set earlier as well as expected running times. Vehicle scheduling assigns sequences of trips from the timetable to specific vehicles, resulting in the sets of trips to be served by each vehicle. Crew scheduling assigns sets of vehicle trips to drivers in accordance with work rules governing shift durations, breaks, and pay provisions. 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. Stochas-

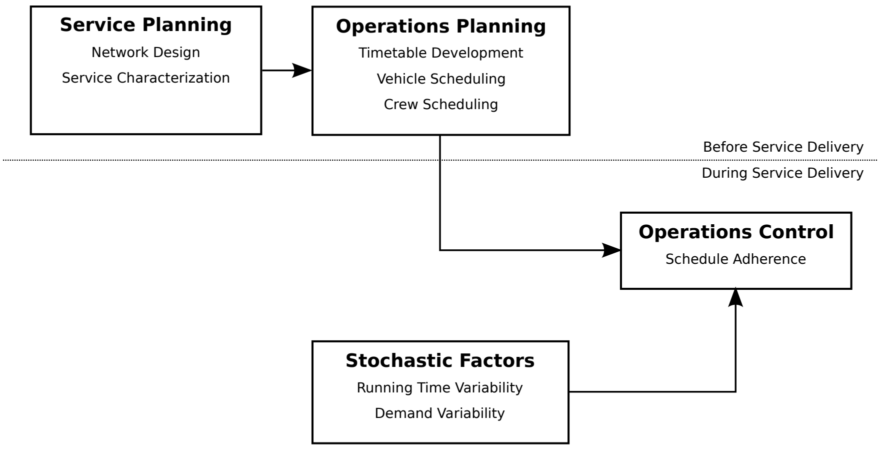

Fig. 1 Schedule-Based Paradigm

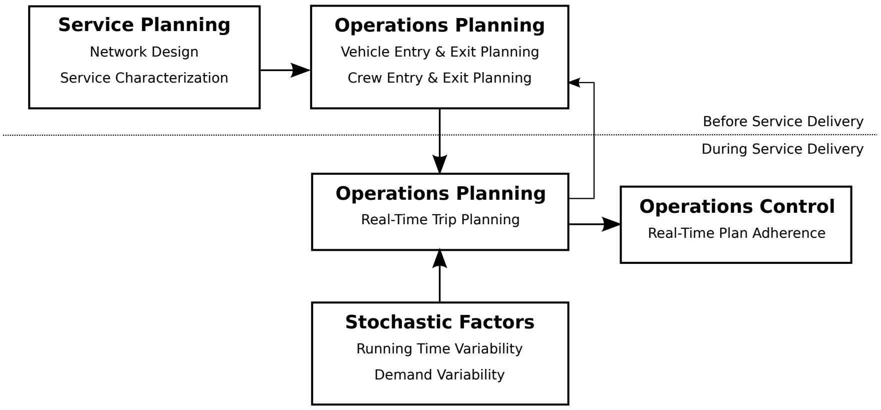

Fig. 2 Schedule-Free Paradigm

tic 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 crews 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 times 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. Strategies employed to meet the operations objective may include holding, short-turning, dead-heading, expressing, and injecting reserve vehicles. 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.

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 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.

A transit system's performance under the schedule-free paradigm feeds back into entry and exit planning. Entry and exit plans limit what can be achieved in real-time, so it can be beneficial to optimize them. 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. 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.

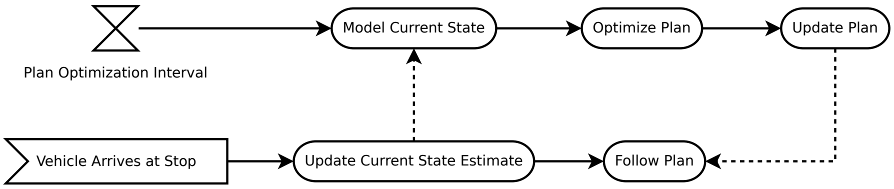

Fig. 3 Schedule-Free Operations Architecture

4 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 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 optimizing *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. 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. The trip plan optimization problem can be formulated as

$$\begin{aligned}
\underset{x\in X}{\text{minimize}} \quad & C(x;p) && (1)\\
\text{subject to} \quad & u_v \le u''_v \quad \forall v \in V && (2)\\
& z_v = z'_v \quad \forall v \in V && (3)\\
& \text{vehicle movement constraints} && (4)\\
& \text{passenger activity constraints} && (5)\\
& h_e^{\min} \le h_e \le h_e^{\max} \quad \forall e \in E && (6)
\end{aligned}$$

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 (2) and (3) ensure that plans deliver vehicles to their exit locations by the required times, while constraint (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:

$$C = C_P + \theta_L C_L + \theta_C C_C \tag{7}$$

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. 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. The discount factor is of the form $\mathrm{e}^{\beta(t-t_0)} \in [0,1]\ ; \beta \le 0$. Mean passenger cost is given by

$$C_P = \frac{\sum_{i=1}^{n} \mathrm{e}^{\beta(t_i-t_0)} \left(V_i + \theta_W W_i\right) + \mathrm{e}^{\beta(t_f-t_0)} \theta_W W_f}{P} \tag{8}$$

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:

$$C_L = \sum_{v \in V} \max\Big(0, (u_v - u'_v)^{\alpha}\Big) \tag{9}$$

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

$$C_L = \sum_{v \in V} \max\Big(0, \mathrm{e}^{\beta(u_v - t_0)}\,(u_v - u'_v)^{\alpha}\Big) \tag{10}$$

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

$$\underset{s \in S}{\text{minimize}} \quad C(s; d^*_s, p) \tag{11}$$

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:

$$\underset{d \in D}{\text{minimize}} \quad C(d; s, p) \tag{12}$$

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 (2) through (6) apply, as before, in both the master problem and the subproblem.

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. 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, but doing so 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.

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

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.

5 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 6. Several simplifications are made in the interest of tractability. Application results presented in Section 6.2 suggest that this methodology does not perform well enough and that refinement is needed. Nevertheless, it lays the groundwork for further exploratory work.

The optimization 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. 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. 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. The remainder of this section discusses each of these steps in greater detail.

Following the discussion in Section 4, the trip sequence problem (11) is decomposed into sequential trip sequence subproblems for each vehicle, as follows:

$$\underset{s_v \in S_v}{\text{minimize}} \quad C(s_v; s_{\bar{v}}, d_s^*, p) \tag{13}$$

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 (13) optimizes the trip sequence of a vehicle v and departure times for all vehicles (through the departure time subproblem (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. The second approach generates the *closest basic sequence*, composed of trips between terminals until the exit closest to the target exit time. In both cases, no holding (beyond the minimum required at terminals in order for drivers to rest between trips) is assumed. Once each vehicle has an initial trip sequence, departure times can be optimized by solving (12) to finish defining an initial solution.

After obtaining an initial solution, a set of feasible trip sequences is generated for each vehicle. The model builds trip sequences 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. 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. Dead-heading is enabled by specifying *dead variations*, which begin and end at turning points but have no stops in between. Limited stop services are enabled by specifying variations that skip stops. The problem's complexity increases exponentially with the number of variations. Feasible trip sequences finish the current trip, do not have trips beginning after u', and exit before u''. 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 (12), and the one resulting in the least cost is selected, thus solving (13). Trip sequences ending after the target exit time u' incur exit lateness cost.

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 (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.

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 with the number of planned events at control points in the horizon. Sánchez-Martínez (2015) 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. A *constrained even headway* policy is applied iteratively to decrease cost, ensuring that holding time constraints are satisfied and vehicles do not exit late. An *event-based performance model* is used to evaluate the cost of each candidate plan. It is based on events representing vehicle arrivals at stops. The reader is referred to Sánchez-Martínez (2015) for further details.

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.

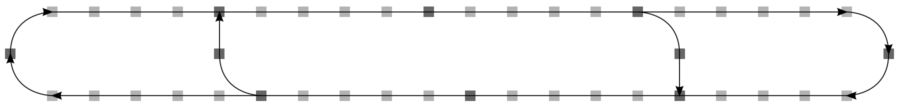

Fig. 4 Simulated Transit Line

6 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 6.1. 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 6.2 compares the performance of the transit line under the schedule-based and schedule-free paradigms with and without delays. Performance is evaluated in terms of passenger cost, i.e. waiting times and in-vehicle time, and driver exit lateness.

6.1 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. Short-turning is allowed at the 15$^{\text{th}}$ stop in each direction (to the 5$^{\text{th}}$ 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. 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 to peak at 3 minutes per link. The schedule assumes no delays. The target headway used to generate the schedule is 8 minutes in the off-peak and 3 minutes in the peak. Demand is modeled as a Poisson process. All origin-destination pairs (in each direction) have the same arrival rate function. The arrival rate in direction 1 is such that vehicle loads reach half the capacity when headways are 8 minutes. The arrival rate 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. 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 (7)) or an exponent α in the lateness cost function (9). Solution complexity cost is neglected, with $\theta_C = 0$ in (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.) 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 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. 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 5. 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.

6.2 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 5.

Table 1 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.

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,

Table 1 Performance Comparison

		Short-Turning	
Delay	Performance Measure	SB	SF
None	Waiting Time (min)	2.6	2.6
	Excess Waiting Time (min)	0.0	0.0
	In-Vehicle Time (min)	9.6	9.5
	Late Exits	0	0
	Max Exit Lateness (min)	0.0	0.0
	Trips	190	192
	Short Turns	0	3
Moderate	Waiting Time (min)	3.3	2.7
	Excess Waiting Time (min)	0.7	0.1
	In-Vehicle Time (min)	10.8	10.7
	Late Exits	0	6
	Max Exit Lateness (min)	0.0	4.1
	Trips	190	190
	Short Turns	2	2

which implies it is driven by a lower predicted passenger cost. By short-turning three times and using the extra 15 minutes, the schedule-free real-time planner manages to serve another cycle.

With delays, 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 can exit late under the schedule-free paradigm 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.

Simulations were run on a computer having an Intel Core i7-3930K processor running at 3.20GHz. With a mean optimization time under 10 seconds and a maximum of 221.2 seconds, it is feasible to plan operations in real-time following the approach presented in Section 5.

7 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.

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.

The schedule-free paradigm is applied to a simulated transit line, with and without delays. 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, in the presence of (unexpected) delays. 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.

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.

References

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

Adenso-Díaz B, González MO, González-Torre P (1999) On-line timetable re-scheduling in regional train services. Transportation Research Part B: Methodological 33(6):387–398, DOI 10.1016/S0191-2615(98)00041-1

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

Bartholdi JJ, Eisenstein DD (2012) A self-coördinating bus route to resist bus bunching. Transportation Research 46B(4):481–491, DOI 10.1016/j.trb.2011.11.001

Boyle DK (2009) TCRP Report 135: Controlling System Costs: Basic and Advanced Scheduling Manuals and Contemporary Issues in Transit Scheduling

Cats O, Larijani AN, Koutsopoulos HN, Burghout W (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

Corman F, D'Ariano A, Pacciarelli D, Pranzo M (2010) A tabu search algorithm for rerouting trains during rail operations. Transportation Research Part B: Methodological 44(1):175–192, DOI 10.1016/j.trb.2009.05.004

Corman F, D'Ariano A, Pacciarelli D, Pranzo M (2012) Bi-objective conflict detection and resolution in railway traffic management. Transportation Research Part C: Emerging Technologies 20(1):79–94, DOI 10.1016/j.trc.2010.09.009

Cortés CE, Jara-Díaz S, Tirachini A (2011) Integrating short turning and deadheading in the optimization of transit services. Transportation Research Part A: Policy and Practice 45(5):419–434, DOI 10.1016/j.tra.2011.02.002

Şahin (1999) Railway traffic control and train scheduling based onintertrain conflict management. Transportation Research Part B: Methodological 33(7):511–534, DOI 10.1016/S0191-2615(99)00004-1

Daganzo CF, Pilachowski J (2011) Reducing Bunching with Bus-to-Bus Cooperation. Transportation Research 45B(1):267–277

D'Ariano A, Pacciarelli D, Pranzo M (2007) A branch and bound algorithm for scheduling trains in a railway network. European Journal of Operational Research 183(2):643–657, DOI 10.1016/j.ejor.2006.10.034

D'Ariano A, Pacciarelli D, Pranzo M (2008) Assessment of flexible timetables in real-time traffic management of a railway bottleneck. Transportation Research Part C: Emerging Technologies 16(2):232–245, DOI 10.1016/j.trc.2007.07.006

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

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

Eberlein XJ, Wilson NH, Bernstein D (2001) The Holding Problem with Real–Time Information Available. Transportation science 35(1):1–18

Huisman D (2007) A column generation approach for the rail crew rescheduling problem. European Journal of Operational Research 180(1):163–173, DOI 10.1016/j.ejor.2006.04.026

Huisman D, Wagelmans AP (2006) A solution approach for dynamic vehicle and crew scheduling. European Journal of Operational Research 172(2):453–471, DOI 10.1016/j.ejor.2004.10.009

Krasemann JT (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, DOI 10.1016/j.trc.2010.12.004

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

10.1016/j.trb.2010.01.003

Mazzarello M, Ottaviani E (2007) A traffic management system for real-time traffic optimisation in railways. Transportation Research Part B: Methodological 41(2):246–274, DOI 10.1016/j.trb.2006.02.005

Mesquita M, Paias A (2008) Set partitioning/covering-based approaches for the integrated vehicle and crew scheduling problem. Computers & Operations Research 35(5):1562–1575, DOI 10.1016/j.cor.2006.09.001

Osuna EE, Newell GF (1972) Control Strategies for an Idealized Public Transportation System. Transportation Science 6(1):52–72

Rezanova NJ, Ryan DM (2010) The train driver recovery problem—A set partitioning based model and solution method. Computers & Operations Research 37(5):845–856, DOI 10.1016/j.cor.2009.03.023

Rodriguez J (2007) A constraint programming model for real-time train scheduling at junctions. Transportation Research Part B: Methodological 41(2):231–245, DOI 10.1016/j.trb.2006.02.006

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

Sánchez-Martínez GE (2015) Real-Time Operations Planning and Control of High-Frequency Transit. PhD thesis, Massachusetts Institute of Technology

Site PD, Filippi F (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, DOI 10.1016/S0965-8564(97)00016-5

Törnquist J, Persson JA (2007) N-tracked railway traffic re-scheduling during disturbances. Transportation Research Part B: Methodological 41(3):342–362, DOI 10.1016/j.trb.2006.06.001

Valouxis C, Housos E (2002) Combined bus and driver scheduling. Computers & Operations Research 29(3):243–259, DOI 10.1016/S0305-0548(00)00067-8

Veelenturf LP, Potthoff D, Huisman D, Kroon LG (2012) Railway crew rescheduling with retiming. Transportation Research Part C: Emerging Technologies 20(1):95–110, DOI 10.1016/j.trc.2010.09.008

Vuchic VR (2005) Urban transit: operations, planning, and economics

Walker CG, Snowdon JN, Ryan DM (2005) Simultaneous disruption recovery of a train timetable and crew roster in real time. Computers & Operations Research 32(8):2077–2094, DOI 10.1016/j.cor.2004.02.001