A cell-based Merchant–Nemhauser model for the system optimum dynamic traffic assignment problem

Yu (Marco) Nie

Department of Civil and Environmental Engineering Northwestern University, 2145 Sheridan Road, Evanston, IL 60208, United States

ARTICLE INFO

Article history:
Received 20 October 2009
Received in revised form 2 July 2010
Accepted 5 July 2010

Keywords:
System optimum dynamic traffic assignment
Cell transmission model
The Merchant–Nemhauser (M-N) model
Marginal cost analysis
Flow holding back

ABSTRACT

A cell-based variant of the Merchant–Nemhauser (M-N) model is proposed for the system optimum (SO) dynamic traffic assignment (DTA) problem. Once linearized and augmented with additional constraints to capture cross-cell interactions, the model becomes a linear program that embeds a relaxed cell transmission model (CTM) to propagate traffic. As a result, we show that CTM-type traffic dynamics can be derived from the original M-N model, when the exit-flow function is properly selected and discretized. The proposed cell-based M-N model has a simple constraint structure and cell network representation because all intersections and cells are treated uniformly. Path marginal costs are defined using a recursive formula that involves a subset of multipliers from the linear program. This definition is then employed to interpret the necessary condition, which is a dynamic extension of the Wardrop’s second principle. An algorithm is presented to solve the flow holding back problem that is known to exist in many discrete SO-DTA models. A numerical experiment is conducted to verify the proposed model and algorithm.

1. Introduction

System optimum (SO) dynamic traffic assignment (DTA) is arguably the most well-studied of all DTA problems. Comparing to its use-equilibrium (UE) counterpart (e.g. Mahmassani and Herman, 1984; Friesz et al., 1993; Smith, 1993; Ran and Boyce, 1996), which seeks dynamic extensions of the Wardrop equilibrium (Wardrop, 1952), system optimum DTA problems enjoy the benefit of having a well-defined objective function (i.e. minimizing total system cost/time) and hence is usually more amenable to analysis. SO-DTA problems have found many applications in transportation planning and operations. The solution of SO-DTA problems has been used as a benchmark to evaluate the benefits of investment decisions (e.g. network expansion) (Waller, 2000; Karoonsoontawong and Waller, 2009), traffic management policies (e.g. congestion pricing and information provision) (May and Milne, 2000; Shen and Zhang, 2009), and operational strategies (e.g. signal control/ramp metering) (Munoz and Laval, 2006). In addition, the past decade has witnessed a concerted effort of applying SO-DTA models in large-scale evacuation planning (e.g. Sattayhatewa and Ran, 2000; Jha et al., 2004; Chin et al., 2005; Chiu et al., 2005; Shyati and Mahmassani, 2006; Liu et al., 2006; Shen et al., 2006).

Few would disagree that the study of SO-DTA problems as known of today was initiated by Merchant and Nemhauser (1978a,b). The Merchant–Nemhauser (M-N) model is a discrete-time network assignment model that seeks to minimize the total system cost of moving flows through a network to reach a single destination, over a pre-determined analysis period. To propagate traffic in a way that reflects congestion effects, the M-N model assumes that the exit flow of each link at a given time interval must be determined by a continuous and concave function, known as the exit-flow function. This assumption

* Tel.: +1 847 467 0502; fax: +1 847 491 4011.
E-mail address: y-nie@northwestern.edu

© 2010 Elsevier Ltd. All rights reserved.
leads to major criticisms that have inspired many to pursue extensions. First, the exit-flow function entails nonconvexity and all associated modeling difficulties because the function is generally nonlinear and has to appear in equality constraints. This problem is often tackled by relaxation (i.e., replacing the nonlinear equality constraints with inequality ones) or linearization. Both remedies could, however, cause the so-called flow holding back problem, essentially because a solution of the relaxed/linearized problem is not always that of the original one. The second and more compelling criticism of the M-N model has to do with its traffic propagation mechanism associated with the use of the exit-flow function. The often cited shortcomings include (e.g., Astarita, 1996; Friesz et al., 2001; Nie and Zhang, 2005): (1) instantaneous flow propagation; (2) violating the anisotropic property of traffic; (3) inability to capture queue spillback effects; and (4) measurement and calibration difficulties. In light of these unsatisfactory properties, exit-flow functions have been considered by many unsuitable to describe traffic dynamics in DTA models (Friesz et al., 2001). Accordingly, several alternatives have been suggested, such as link delay (or exit time) functions (Friesz et al., 1993; Carey and Subrahmanian, 2000), variants of the bottleneck model (Chang et al., 1988; Smith, 1993), and the celebrated cell transmission model (CTM) (Daganzo, 1994, 1995; Ziliaskopoulos, 2000). However, it has been pointed out (Carey and McCartney, 2004) that some drawbacks attributed to exit-flow functions (such as 1, 2 and 4 listed above) can be resolved so long as one is willing to further discretize space. Also, the spillback effects could be addressed, either by introducing a downward sloping portion into the exit-flow function, or by explicitly restricting flow according to downstream conditions (Carey and McCartney, 2004). In part, these observations motivated our study.

The first and primary goal of this paper is to reveal that the CTM-type traffic dynamics can be derived from the original M-N model. Specifically, when a special piecewise linear exit-flow function is adopted, the discretized M-N model would encapsulate point-queue traffic dynamics as its flow propagation scheme. If additional constraints are added to respect downstream restrictions, it effectively replicates the CTM-based SO-DTA model of Ziliaskopoulos (2000).

Secondly, from the above finding stems a revised M-N model, which represents traffic propagation with the same fidelity as the existing CTM-based SO-DTA models, but has a simpler constraint structure and requires a simpler cell network representation. This structural simplicity is derived from the fact that the revised M-N model treats all intersections and cells uniformly. In contrast, the existing CTM-based models typically require decomposing general intersections into standard merge and diverge, which may increase the size of the resulting problem and introduce operational complexity into the modeling process.

Thirdly, linking the revised M-N model to the CTM-based model suggests that the flow holding back problem, which is known to exist in many discrete SO-DTA models (e.g., Ziliaskopoulos, 2000; Carey and Subrahmanian, 2000; Shen et al., 2006), is rooted in the violation of the ordered solution property (OSP) imposed in the linearization process of the M-N model. Accordingly, we show that the algorithm of Ho (1980), initially developed for the M-N model, can be tailored to treat this daunting problem.

Finally, a marginal cost analysis conducted in this study offers a recursive formula to evaluate the path marginal cost. The validity of the formula is established by proving that it satisfies the necessary condition implied by the Wardrop's second principle, namely, all used paths at SO should have the same and minimum marginal cost as defined by the recursive formula. Among other utilities, the marginal cost analysis enables a potentially more efficient solution approach that solves the SO-DTA problems by equilibrating marginal path costs.

Section 2 briefly reviews the SO-DTA studies in order to better position our work in the literature. Section 3 presents the cell-based M-N model and reveals its relationship to existing models. The marginal cost analysis and the discussion of the flow holding back problem are presented in Sections 4 and 5, respectively. Section 6 gives a numerical example. Section 7 concludes the study.

2. Literature review

Existing SO-DTA models can be classified into discrete- and continuous-time models, according to primarily whether or not the traffic flow propagation is discretized with respect to time. We exclude from our review “simulation-based” models, which are constructed on the premise that the SO-DTA problem can be solved as a UE problem. These models often evaluate path marginal costs using traffic simulation or dynamic network loading (e.g., Chali and Smith, 1995; Peeta and Mahmassani, 1995; Shen et al., 2007).

2.1. Discrete-time models

Most discrete SO-DTA models share a similar mathematical programming structure with solution variables represented by link volumes and entry/exit flows, an objective function that represents the total system cost (usually travel times), and constraints capturing flow conservation and traffic propagation. To the best of our knowledge, the model of Merchant and Nemhauser (1978a,b) is the first of this kind, which we shall revisit in details in the following sections. Carey (1987) proposed a convex reformulation for the M-N model, by replacing the exit-flow function with a control variable and relaxing the nonlinear equality constraints. Marginal costs and externalities of the convex model and their implications in congestion pricing are discussed in Carey and Srinivasan (1993). Carey (1992a) conducted a marginal cost analysis for a SO-DTA model based on the point-queue traffic flow model. The focus of Carey (1992a) is to analyze the marginal cost induced by the first-in-first-out (FIFO) constraint, which becomes necessary when multi-commodity flows (such as flows heading to multiple
destinations) is concerned. However, neither Carey (1992a) nor Carey and Srinivasan (1993) considers queue spillback effects.

The model of Chang et al. (1988) is built on a space-time expansion network (STEN), which is constructed to represent the point-queue model. Both route and departure time choices are considered in their study. The similar idea has been pursued by Yang and Meng (1998) and Carey and Subrahmanian (2000), while the latter discretizes a delay function model using the STEN concept. In general, the objective of a STEN-based approach is to create a static equivalence of the dynamic problem. While the structure of the expanded network may be exploited to improve solution efficiency (for a recent example cf. Shen et al., 2006), explicitly constructing STEN generally incur significant computation overhead in large problems.

The cell transmission model (Daganzo, 1994, 1995) was first employed to tackle SO-DTA problems by Ziliaskopoulos (2000). He formulated the singe-destination SO-DTA problem as a standard linear program encapsulating relaxed CTM traffic dynamics. Ziliaskopoulos’s model has since been widely used thanks to its simple linear structure and ability to capture queue spillback. Chiu and Zheng (2007) extended the CTM-based SO-DTA model to accommodate multiple destinations by regulating prioritized flows through permissible destinations. However, their model does not treat the flow holding back problem and the potential FIFO violations.

2.2. Continuous-time models

Friesz et al. (1989) reformulated the M-N model as a continuous optimal control problem and interpreted the first-order condition using a so-called instantaneous path marginal cost. An iterative solution method for the continuous model was proposed in Wie et al. (1994). The outer loop of that method penalizes the equality constraints using an augmented Lagrangian algorithm, while the inner loop solves the optimal control subproblem. The model of Wie and Tobin (1998) aims to remedy the instantaneous flow propagation imposed by the exit-flow function in the continuous, link-based M-N model. Their model divides each link into a free-flow section and a queueing section, and only applies the exit-flow function to the latter. Recently, Chow (2007, 2009) reformulated the continuous M-N model using the link delay function of Friesz et al. (1993). Chow (2009) also proposed a solution algorithm along the line of Wie et al. (1994).

2.3. Summary

There are issues that neither continuous nor discrete models have hitherto effectively addressed. For example, enforcing the FIFO condition to avoid “forced” over-taking in favor of the system cost is known to introduce nonconvexity Carey (1992b). Hence, the applicability of most SO-DTA models is restricted in theory to the single-destination single-commodity problem. Continuous and discrete models also have their own advantages and disadvantages. Continuous models promise analytical insights (such as the closed form externality analysis) and greater computational tractability. Nevertheless, cross-link interactions such as queue spillback effects are difficult to capture within the existing framework of continuous models. Due to the complex model structure, they are generally less amenable to efficient solution procedures, even though they typically bypass direct network expansion. Indeed, solving a continuous model on a real-sized problem has rarely been attempted. On the other hand, linearized discrete models, such as the CTM-based model, effectively circumvents the analytical challenges entailed by complex flow propagation, and leads to simple linear programs. However, discrete models also compromise computational tractability since they often lead to very large and sparse LP equivalence. Another daunting problem shared by many discrete models is the flow holding back problem stemming from relaxation and linearization (e.g. Merchant and Nemhauser, 1978a; Carey and Subrahmanian, 2000; Ziliaskopoulos, 2000).

The model discussed in what follows is a variant of the M-N model with a special exit-flow function and extra constraints. It is equivalent to the CTM-based SO-DTA model but has a simpler constraint structure. Also, our model is amenable to a known resolution to the flow holding back problem.

3. A cell-based M-N formulation

Consider a directed network $G(N, A)$, where $N$ and $A$ are the sets of nodes and links respectively. Let $R$ represent the set of origins and $s$ the single destination. The assignment horizon is divided into a finite number of discrete time intervals indexed by $\{t : t = 0, 1, \ldots, T\}$. To better describe traffic dynamics, each link is divided into $\left\lfloor \frac{d}{\delta t} \right\rfloor$ cells, where $\Delta$ is the uniform length of discrete time intervals, $\delta t$ is free flow speed on link $I$, $d$ is the length of link $I$, and $\lfloor w \rfloor$ denote the largest integer no greater than $w$. We thus obtain a cell-based representation of the original network denoted as $G(N', A')$, where $A'$ is the set of cells as defined above, $N'$ is augmented from $N$ by the intra-link cell junctions. Note that the transformation neither requires simplifying the original intersection layout nor affects the set $R$ and the destination $s$. Indeed, this cell-based network is no different than a conventional, link-based network, except that by construction the traversal time on any of its links is one (note that a “link” on the cell-based network is actually a cell).

1 We note that the definition of the marginal path cost in Merchant and Nemhauser (1978b) is not an instantaneous one; neither is our definition presented in Section 4.

2 Such over-taking could take place among vehicles with different cost functions or heading to different destinations.
Let $O(i)$ and $I(i)$ be the sets of outgoing and incoming cells of node $i \in N_f$, respectively. Let $q_{it}^a$ be the travel demand between O–D pair $rs$ departing at time $t$, and the numbers of vehicles on, entering, and exiting a cell $a$ at time $t$ are denoted by $x_t^a$, $u_t^a$, $v_t^a$, respectively.

Following the original M-N model, we assume that the number of vehicles exiting cell $a$ during time interval $t$, $v_t^a$, is uniquely determined by the number of vehicles on that cell at $t$, $x_t^a$. Namely, $v_t^a = g^a(x_t^a)$. To be physically meaningful, the function $g^a(\cdot)$ should be nondecreasing, continuous and concave. Note that by definition

$$g^a(x) \leq x - g^a(x) \leq 1$$

The cost of traversing a cell, denoted as $h^a(x^a)$, is also uniquely determined by $x^a$. The total system cost can then be represented by summing up $h^a(x^a)$ for all links over all the time intervals. The cost function $h$ is generally assumed to be nonnegative, nondecreasing, and convex. We are now ready to present the cell-based M-N model as follows:

$$\min_{\{x, u, v\}} \sum_{t=1}^{T} \sum_{a \in A} h^a(x_t^a)$$

s.t. $x_{t+1}^a = x_t^a - g^a(x_t^a) + u_t^a, \quad t = 0, \ldots, T - 1, \quad \forall a \in A^c$ (2b)

$u_t^a \geq 0, \quad t = 0, \ldots, T - 1, \quad \forall a \in A^c$ (2c)

$x_t^a \geq 0, \quad t = 1, \ldots, T, \quad \forall a \in A^c$ (2d)

Constraint (2b) describes traffic flow propagation; Constraint (2c) represents node flow conservation; Constraint (2d) is the initial condition while (2e) and (2f) are nonnegative constraints. We note that, because of (1), Constraints (2b), (2d) and (2e) imply (2f).

The M-N model is a linear program (LP) if both $h^a(\cdot)$ and $g^a(\cdot)$ are linear functions. Modeling $h^a$ as a linear function seems acceptable in many cases; for instance $h^a(x^a) = x^a$ implies that the total system cost is the total travel time, which is indeed what one often seeks to minimize in reality. However, it was deemed necessary to introduce some form of nonlinearity to $g^a(\cdot)$ in order to capture congestion effects properly. It is this consideration that complicates the matter. First, once a nonlinear term enters equality constraints, the problem becomes nonconvex. Second, the existence of these nonlinear constraints raises the question about whether the KKT conditions are indeed necessary for optimality – recall that this holds only when the constraint structure satisfies certain qualification conditions. The latter issue was resolved by Carey (1986), who showed that the M-N model satisfies the linear independent constraint qualification (LICQ). The issue of nonconvexity is tackled in Merchant and Nemhauser (1978a) through linearization. For each $x^a$, the nonnegative segment of the real line is partitioned into $K(a)$ subintervals $[l_k^a, u_k^a]$, where $l_k^a = 0, u_k^a = \infty$ and $l_k^a = u_{k+1}^a$ for $k = 2, \ldots, K(a)$. During each subinterval $[l_k^a, u_k^a]$, both $g^a(\cdot)$ and $h^a(\cdot)$ are approximated by affine functions with slopes $g_k^a = [g^a(u_k^a) - g^a(l_k^a)]/(u_k^a - l_k^a)$, and $h_k^a = [h^a(u_k^a) - h^a(l_k^a)]/(u_k^a - l_k^a)$, respectively. Corresponding to each subinterval, variables $\{x_{ik}^a\}$ are introduced so that $0 \leq x_{ik}^a \leq C_k = u_k^a - l_k^a$. See Fig. 1 for an illustration of the linearization scheme when $K(a) = 4$.

![Fig. 1. Illustration of piecewise linearization of $h^a$ and $g^a$ when $K(a) = 4$.](image-url)
Now, we can replace $x_t^a$ with $\sum_{k=1}^{K(a)} x_t^{a_k} g^a(x_t^{a_k})$ with $\sum_{k=1}^{K(a)} x_t^{a_k}$, and $h^a(x_t^a)$ with $\sum_{k=1}^{K(a)} h^a_k x_t^{a_k}$. The linearized M-N model reads:

$$\min_{\{x_t^{a_k}, \alpha_t\}} \sum_{t=1}^{T} \sum_{a \in A} \sum_{k=1}^{K(a)} h^a_k x_t^{a_k}$$

subject to

$$\sum_{k=1}^{K(a)} x_{t+1}^{a_k} = \sum_{k=1}^{K(a)} (1 - g^a_k) x_t^{a_k} + \alpha_t, \quad t = 0, \ldots, T - 1, \quad \forall \ a \in A^c$$

$$\sum_{a \in \Omega(i)} \alpha_t = q_t + \sum_{a \in \Omega(i)} x_t^{a_1}, \quad t = 0, \ldots, T - 1, \quad \forall \ i \in N', \ i \neq s$$

$$\sum_{k=1}^{K(a)} x_t^{a_k} = 0 \text{ (given)}, \quad \forall \ a \in A^c$$

$$\alpha_t \geq 0, \quad t = 0, \ldots, T - 1, \quad \forall \ a \in A^c$$

$$0 \leq x_t^{a_k} \leq C_t^a, \quad t = 1, \ldots, T, \quad \forall \ a \in A^c, \ k = 1, \ldots, K(a)$$

The linearized model (3) is not a standard LP because of the new constraint (3g), which, known as the ordered solution property (OSP) constraint, requires that the piecewise linear solution meaningfully represents the exit and cost functions. However, Merchant and Nemhauser (1978a) proved that if $h^a(\cdot)$ satisfies the so-called linear cost function assumption (LCFA), there must exist an optimal solution to the LP (3a)–(3e) and (3f) that satisfies the OSP constraints. The reader is referred to Merchant and Nemhauser (1978a) for the details of LCFA and the proof. Suffice it to mention here that LCFA is always satisfied for the following two cost functions:

1. $h^a(x_t^a) = x_t^a$; 2. $h^a(x_t^a) = \left\{ \begin{array}{ll} x_t^a & \text{if } t = T \\ 0 & \text{otherwise} \end{array} \right.$

The first function sets the system cost as the total travel time and will be considered hereafter as our objective function. Nevertheless, a solution obtained from a standard LP algorithm such as the simplex method may well violate OSP constraints. This violation is precisely the reason why the flow holding back problem arises in many discrete models, including the CTM-based model, as explained later.

We now construct a special instance of the model (3) by selecting the following exit-flow function:

$$g^a(x_t^{a_k}) = \left\{ \begin{array}{ll} x_t^{a_k} & x_t^{a_k} \leq C_t^a \\ C_t^a & \text{otherwise} \end{array} \right.$$

In this case, the cost function is an affine function with a slope of one. The exit-flow function consists of two linear segments: the first segment has a slope of one and the second has a slope of zero. Thus, using the same discretization scheme as shown in Fig. 1, these functions can be represented using two piecewise linear segments, i.e. $K(a) = 2$. Note that after discretization $g_t^1 = 1$ and $g_t^2 = 0$ for the selected exit function, and that $h_t^1 - h_t^2 = 1$ for the selected cost function. Thus, the model (3) can be revised as follows:

$$\min_{\{x_t^{a_k}, \alpha_t\}} \sum_{t=1}^{T} \sum_{a \in A} \sum_{k=1}^{2} x_t^{a_k}$$

subject to

$$\sum_{k=1}^{2} x_{t+1}^{a_k} = \sum_{k=1}^{2} (1 - g^a_k) x_t^{a_k} + \alpha_t, \quad t = 0, \ldots, T - 1, \quad \forall \ a \in A^c$$

$$\sum_{a \in \Omega(i)} \alpha_t = q_t + \sum_{a \in \Omega(i)} x_t^{a_1}, \quad t = 0, \ldots, T - 1, \quad \forall \ i \in N', \ i \neq s$$

$$\sum_{k=1}^{2} x_t^{a_k} = 0 \text{ (given)}, \quad \forall \ a \in A^c$$

$$\alpha_t \geq 0, \quad t = 0, \ldots, T - 1, \quad \forall \ a \in A^c$$

$$0 \leq x_t^{a_k} \leq C_t^a, \quad t = 1, \ldots, T, \quad \forall \ a \in A^c, \ k = 1, \ldots, K(a)$$

Let us now introduce $x_t^{a_{12}} \equiv p_t^a$ and $x_t^{a_{11}} \equiv v_t^a$, and interpret the new variable $p_t^a$ as the queued flows in cell $a$ at the end of time $t$, and $v_t^a$ as the exit flow from cell $a$ during time $t$. We emphasize that, with the above definition, $u_t^a$ should be understood as the inflow rate into cell $a$ at the end of time $t$, or at the beginning of time $t + 1$. The model (5) can then be written as:

$$\min_{\{p_t^a, v_t^a\}} \sum_{t=1}^{T} \sum_{a \in A} (p_t^a + v_t^a)$$

subject to

$$\sum_{a \in \Omega(i)} p_{t+1}^a = q_t + \sum_{a \in \Omega(i)} v_t^a, \quad t = 0, \ldots, T - 1, \quad \forall \ a \in A^c$$

$$\sum_{a \in \Omega(i)} v_t^a = q_t + \sum_{a \in \Omega(i)} p_t^a, \quad t = 0, \ldots, T - 1, \quad \forall \ i \in N', \ i \neq s$$
\[ v_0^a + p_0^a = R^a \geq 0 \text{ (given)} \quad \forall a \in A^c \] (6d)

\[ u_t^a \geq 0, \quad t = 0, \ldots, T - 1, \quad \forall a \in A^c \] (6e)

\[ 0 \leq v_t^a \leq C^a, \quad t = 1, \ldots, T, \quad \forall a \in A^c \] (6f)

\[ p_t^a > 0 \Rightarrow v_t^a = C^a, \quad t = 1, \ldots, T, \quad \forall a \in A^c \] (6g)

Note that the nonnegative condition \( p_t^a \geq 0 \) is not included in (6) because it is implied from (1) \( g^a(x_t^a) = v_t^a \leq C^a = v^f(t) + p_t^a(t) \Rightarrow p_t^a(t) \geq 0 \). Interestingly, this model embeds the point-queue traffic dynamics, although it is linearized from an exit-flow function model. To see this, first note that (6f) states that the exit flow should be bounded by \( C^a \) defined in (4), which can be safely interpreted as the "capacity" of the cell \( a \). Second, Eq. (6b) states that the queued flows at time \( t \) is the sum of the queued and incoming flows from the last interval \((t - 1)\) subtracting the exit flow at \( t \). Recall in a continuous point-queue model (e.g. Kuwahara and Akamatsu, 1997), the derivative of queuing \( dp \) with respect to \( t \) is

\[ \frac{dp}{dt} = u(t - \tau_0) - v(t) \] (7)

where \( u(t) \) and \( v(t) \) are entry and exit flow rates, and \( \tau_0 \) is the free flow travel time. One can verify that (6b) is a discrete version of (7) when \( \tau_0 = 1 \). Finally, the OSP constraint (6g) requires the cell to be operated at capacity at \( t \) when a queue is present at the end of \( t \). Violating this condition thus implies "holding" flows when there is still spare capacity.

We now address cross-cell interactions in order to capture queue spillback. This can be achieved by forcing the inflow rate \( u_t^a \) to respect the capacity and "supply" of cell \( a \). The capacity constraint requires

\[ u_t^a \leq C^a, \quad t = 0, 1, \ldots, T - 1 \] (8)

The supply constraint, in the framework of CTM (Daganzo, 1995),\(^3\) can be written as

\[ u_t^a \leq \delta(H^a - p_t^a - v_t^a) \Rightarrow u_t^a + \delta(p_t^a + v_t^a) \leq \delta H^a, \quad t = 0, 1, \ldots, T - 1 \] (9)

where \( H^a \) is the maximum amount of flow that cell \( a \) can hold (often known as the holding capacity), and \( \delta \) is a parameter related to road properties, which can be interpreted as the ratio of the wave speed to the free flow speed. Note that \( 0 < \delta \leq 1 \). Constraint (8) states that the inflow rate to a cell cannot exceed the capacity of that cell; and Constraint (9) states that the inflow rate cannot exceed an upper bound dictated by the holding capacity. The key is to recognize that, since \( u_t^a \) interacts with the exit flow of any upstream cells through the flow conservation condition (6c), limiting \( u_t^a \) by these downstream restrictions ensures that the corresponding congestion effects will be propagated through cells and intersections. Consequently, the spillback effects can be captured.

With the introduction of these additional restrictions, however, the OSP constraint has to be redefined. Essentially, the flow holding occurs only if (1) the queue is present and (2) all three restrictions are not binding. The redefined OSP constraint thus reads (see Shen et al., 2006 for a similar definition):

\[ p_t^a > 0 \Rightarrow v_t^a = C^a, \quad \text{or} \quad u_t^a = C^a \quad \forall b \in O(a^+), \quad \text{or} \quad u_t^a + \delta(p_t^b + v_t^b) = \delta H^b, \quad \forall b \in O(a^+) \] (10)

where \( a^+ \) (\( a^+ \)) denotes the downstream (upstream) node of cell \( a \). We now finalize the proposed cell-based model as follows:

\[
\begin{align*}
\min_{\{v_t^a, p_t^a\}} & \sum_{t=1}^{T} \sum_{a \in A} (p_t^a + v_t^a) \tag{11a} \\
\text{s.t.} \text{[Original constraints]} & \text{(6b–6f)} \tag{11b} \\
& \text{[Cross-cell interactions]} \text{ (8) and (9)} \tag{11c} \\
& \text{[OSP constraint]} \text{ (10)} \tag{11d}
\end{align*}
\]

We note that the model defined by (11a, 11b and 11c) is consistent with the relaxed CTM traffic dynamics as used in Ziliaskopoulos (2000). In particular, the cross-cell influx used in Ziliaskopoulos (2000) can be represented by either \( u_t^a \), \( v_t^a \), or both in our model, depending on the type of cells considered therein (diverge, merge, or ordinary cells). However, (11) does not require decomposing general intersections into standard merge and diverge, which, besides operational inconvenience, could significantly increase the problem size.

\(^3\) Ziliaskopoulos (2000) explains in detail why the CTM traffic dynamics can be relaxed using a set of inequalities similar to (8) and (9). We call it a "relaxation" because the relaxation replaces a min operator \( x = \min(a, b, c) \) with inequalities \( x \leq a, x \leq b, x \leq c \). It is possible, however, that a solution to the relaxed problem is not one for the original problem, i.e., the relaxed LP may end up with a solution \( x < \min(a, b, c) \). The OSP constraints are needed to force the relaxed LP to forgo such solutions. We also note that the merge and diverge ratios present in the original CTM model are not needed here as they will be determined endogenously by the SO-DTA model.
4. Marginal cost analysis

We proceed to derive the KKT conditions for the LP problem (11), in order to propose a formula to evaluate path marginal costs. Note that all discussions presented in this section ignore the OSP constraint. We shall show how these nasty constraints can be dealt with in Section 5.

Associate the multiplier $\lambda_k$ with Constraint (6b), $\mu_t^a$ with (6c), $\alpha_t^a$ with (6f) (upper bound), $\beta_t^a$ with (8), and $\gamma_t^a$ with (9). The optimality conditions are

\[
\begin{align*}
\text{(12a)} & \quad 1 - \lambda_{t-1}^a - \mu_t^a + \alpha_t^a + \delta \gamma_t^a \geq 0 \quad \forall \ a, \ j = a^+, \ j \neq s, \ t = 1, \ldots, T - 1 \\
\text{(12b)} & \quad 1 - \lambda_{t-1}^a + \alpha_t^a + \delta \gamma_t^a = 0 \quad \forall \ a, \ t = 1, \ldots, T - 1 \\
\text{(12c)} & \quad 1 - \lambda_t^a + \alpha_t^a \geq 0 \quad \forall t, \ a \in I(i), \ i \neq s, \ t = 0, \ldots, T - 1 \\
\text{(12d)} & \quad 1 - \lambda_{T-1}^a = 0 \quad \forall a \\
\text{(12e)} & \quad 1 - \lambda_{t-1}^a + \alpha_t^a = 0 \\
\end{align*}
\]

We note that the complementarity structure in (12a), (12c) and (12d) comes from nonnegativity of the associated variables. However, the nonnegativity of $\mu_t^a$ is implied from other conditions and hence not explicitly considered. Plugging Conditions (12b) and (12e) into (12a) and (12d), respectively, we obtain

\[
\begin{align*}
\text{(13a)} & \quad \mu_t^a - \lambda_t^a + \alpha_t^a \geq 0 \quad \forall \ a, \ j = a^+, \ j \neq s, \ t = 1, \ldots, T - 1 \\
\text{(13b)} & \quad \alpha_t^a = 0 \\
\end{align*}
\]

The complementarity conditions require

\[
\begin{align*}
\text{(14a)} & \quad \nu_t^a \leq C^a \\
\text{(14b)} & \quad \alpha_t^a \leq C^a \\
\text{(14c)} & \quad \beta_t^a \leq C^a \\
\text{(14d)} & \quad \gamma_t^a \leq 0 \\
\end{align*}
\]

Consider a path $k$ that consists of a sequence of cells $a_0, \ldots, a_m$, written as $k = (a_0, \ldots, a_m)$. Without loss of generality, we assume that $a_m = s$. The path $k$ can also be written as $\{a_0, k\}$, where $k$ denotes a subpath of $k$ starting from cell $a_0$. We propose to define the marginal cost of this path at time $t$, denoted as $\phi_t^k$, using the following recursive relationship

\[
\phi_{t-1}^k = 1 + \delta \alpha_0 \beta_t^0 + \alpha_0^0 + \phi_t^0 + \beta_t^0 + \gamma_t^0, \quad \forall t = 1, \ldots, T \\
\phi_T^k = 0; \quad \phi_{t-1}^0 = 0, \quad \forall t = 0, \ldots, T
\]

As explained in the following, at optimum, the path marginal cost defined above refers to the system cost increase corresponding to a unit of flow added into its first cell. Fig. 2 demonstrates how $\phi_t^k$ can be constructed for a given path following a decreasing order of time, starting from $T - 1$.

Fig. 2. Illustration of path marginal cost evaluation ($k_1, k_2, k_3$ denotes paths starting at $a_1, a_2, a_3$, respectively).
Theorem 1. If vectors $p$, $v$, $u$ solve Problem (11), then the following conditions must be satisfied for any $t = 0, \ldots, T$, $i \in N^t$, $i \neq s$, and $k = (a_0, \ldots, a_m)$:

1. $\phi^k_t \geq \lambda^k_{0i}$;
2. $\phi^k_{t+1} = \lambda^k_{0i}$ if $u^k_{0i} > 0$, $v^k_{0i+1} > 0$ for $(w, j) = (t + 1, j), (t + 2, j + 1), \ldots$, where $w \leq T - 1$, $j \leq m$.

Proof. We prove the result by induction, following the approach taken by Merchant and Nemhauser (1978b). For $t = T - 1$, we note that (15) becomes

$$\phi^k_{T-1} = 1 + x^k_{0T} + \phi^k_T = 1 + x^k_{0T}$$

where the first equality holds because constraints (8) and (9) do not exist at $T$ (so the associated multipliers diminish); the second equality follows directly from the definition of the marginal cost. Recalling (12d), we have $\phi^k_{T-1} = 1 + x^k_{0T} \geq \lambda^k_{0T-1}$, which is Condition 1. Also from the complementarity of (12d), we know that $\phi^k_{T-1} = \lambda^k_{0T-1}$ when $u^k_{0T} > 0$. Thus, the statement is true at $T - 1$.

Now suppose that the statement is true for time intervals $t, \ldots, T - 1$, $\forall k \in K^t$, $\forall i \neq s$. Let $j = a_0$ and we have

$$\phi^k_{t+1} = 1 + \delta^{o_0, o_{a_0}} + x^k_{0t} + \phi^k_t + \rho^k_t + \gamma^k_t \geq 1 + \delta^{o_0, o_{a_0}} + x^k_{0t} + \phi^k_t - \lambda^k_{0t} + \mu^k_t \geq 1 + \delta^{o_0, o_{a_0}} + \phi^k_t + \lambda^k_{0t} - \lambda^k_{0t}$$

The first inequality follows from (12c); the second inequality follows from (13a); the third inequality follows from the induction assumption; and finally the equality follows from (12b). We thus prove Condition 1 at $t - 1$. To prove Condition 2, we use a similar argument but instead observe that when $u^k_{0t} > 0$ and $v^k_{0t} > 0$,

$$\phi^k_{t-1} = 1 + \delta^{o_0, o_{a_0}} + x^k_{0t} + \phi^k_t - \lambda^k_{0t} + \mu^k_t$$

Finally, from the induction we know that $u^k_{0t} > 0$, $v^k_{0t+1} > 0$ for $(w, j) = (t + 1, j), (t + 2, j + 1), \ldots$, where $w \leq T - 1$, $j \leq m$. Thus for $w = t - 1$, we now know $u^k_{0t} > 0$, $v^k_{0t} > 0$, which completes our proof.

Note that $\lambda^k_{0t}$ represents the marginal cost of adding one unit of flow into cell $a$ at time $t$. Theorem 1 establishes that $\phi^k_t$ given in Eqs. (15), (16) is a proper representation of path marginal costs, because $\phi^k_t$ attains $\lambda^k_{0t}$ when the path is used starting from time $t$ onward, and is always larger than or equal to it otherwise. The path marginal cost can also be interpreted using $\mu^k_t$, which corresponds to the cost of adding one unit of flow at node $i = a_0$, where $a_0$ is the first link of the path. Condition (12c) implies

$$\mu^k_t \leq x^k_{0t} + \rho^k_t + \gamma^k_t, \quad a \in O(i)$$

Furthermore, when $u^k_{0t} > 0$, $a_0 \in O(i)$, we have

$$\mu^k_t = \lambda^k_{0t} + \rho^k_t + \gamma^k_t = \phi^k_t + \rho^k_t + \gamma^k_t, \quad k = (a_0, k)$$

One might argue that the above discussion is of limited use for the purpose of computation, since the evaluation of marginal costs relies on the knowledge of multipliers, which are typically unknown until the problem is solved. However, we note that the analysis enables one to evaluate path marginal costs from multipliers $\phi^k_t$, $\rho^k_t$, and $\gamma^k_t$, which could be iteratively constructed or estimated from the current primal solution. For example, $\phi^k_t$ can be evaluated for a given flow pattern when downstream restriction is negligible; see (Shen et al., 2007) for a detailed discussion. In that case, the proposed recursive formula (15), (16) can be used to compute path marginal costs, which effectively converts the SO assignment to an equivalent UE assignment. How to compute $\rho^k_t$ and $\gamma^k_t$ in a similar fashion so as to take advantages of the marginal cost analysis is a subject of future investigation.

5. Resolving flow holding back

The cell-based linear program (11) can be solved by standard techniques such as the simplex method. However, an optimal solution obtained from such solution methods may not satisfy the OSP constraint (10) and hence could lead to the flow holding back problem. Note that this paper is not concerned with the circumstances where holding is beneficial to the system, such as those identified and discussed in Ziliaskopoulos (2000). Rather, we aim to resolve unnecessary holdings, which can be avoided without increasing the system cost. Removing unnecessary holdings is more than just a mathematical exercise. Indeed, a holding-free SO-DTA solution is preferable because it is less expensive to implement than those with artificial holdings (Shen et al., 2006).

Merchant and Nemhauser (1978a) showed that a holding-free SO solution always exists for the linearized model (3). Recently, this result is found to hold even when cross-cell interactions are considered (Shen et al., 2006). Shen et al. (2006) also proposed a network simplex method which is able to identify and remove holdings in the SO-DTA model based on the
point-queue model. How to remove such holdings in a cell-based model such as (11) and the model of Ziliaskopoulos (2000), however, has not received much attention.

Section 3 showed that the cell-based model is in effect equivalent to the linearized M-N model augmented by additional constraints to capture cross-cell interactions. Since those additional constraints do not change the underlying model properties, existing algorithms for the M-N model, such as Ho’s algorithm Ho (1980), also apply to the cell-based model. Ho’s algorithm overcomes the OSP violations through a successive readjustment scheme. The idea is to fix the violation sequentially following a decreasing order of time. The algorithm starts by solving the original model to get the optimal system objective value, and then solves a sequence of optimization problems to maximize total exit flows, while forcing the solution to satisfy the original constraints and additional “optimality cuts”. We show in the following how Ho’s algorithm can be adapted to resolve the flow holding back problem in our model.

Define $z_r$ as the optimal solution to the LP model (11), without considering the OSP constraint (10). Let $S_r$ be the set of optimal solutions to the original problem, and $S_r-1$ be a subset of $S_r$ that maximizes the total exit flows up to $T - 1$, denoted as $E_r-1 = \sum_{t=0}^{T-1} v_t$. If there are holding back problems at $T - 1$, the flows that are scheduled to leave at time $T$ in the original solution will be moved back to $T - 1$ to maximize the exit flow at $T - 1$, because the solution variables at $T$ are not included in the objective function of the maximization problem. In order that the total system cost remains unchanged in that process, a constraint should be added into the maximization problem. One can perform the above procedure sequentially, while continuously reducing the size of $S_r$ as $t$ changes from $T$ down to 1, where all possible OSP violations should have been overcome. Although Ho’s algorithm requires solving the LP problems for $T$ times in theory, a significant number of iterations is needed only when network is “extremely overloaded” (Ho, 1980). Also, the solutions obtained from the previous problems can be used as a warm start to solve subsequent problems, which could substantially improve the overall efficiency. Finally, specialized algorithms can be designed to take advantage of the network structure. The algorithm is given below; see Ho (1980) for a formal proof of convergence.

**Ho’s algorithm for cell-based M-N model**

- **Step 0:** Solve the LP problem (11) to get the optimal system cost $z_r$. Set $w = T$.
- **Step 1:** Test the OSP constraint (10), stop if satisfied; otherwise, initialize the set of constraints as

$$\Omega = \left\{ p_t^a, v_t^a, \quad \forall \; a, t(11b)-(11c) \quad \text{and} \quad \sum_{t=1}^{T} \sum_{a \in A} (p_t^a + v_t^a) = z_r \right\}$$

- **Step 2:** Solve the following optimization problem.

$$\max_{E_{w-1}} \sum_{t=1}^{w} \sum_{a \in A} v_t^a$$

Denote the optimal objective function value as $E_{w-1}$ and then go to Step 3.
- **Step 3:** Set $w = w - 1$; test the OSP constraint (10), stop if satisfied; otherwise, add the following constraint into $\Omega$:

$$E_w = \sum_{t=1}^{w} \sum_{a \in A} v_t^a = E_{w-1}$$

and then go back to Step 2.

6. Numerical example

For the purpose of demonstration and comparison, we implement the cell-based M-N model and solve the benchmark problem given in Ziliaskopoulos (2000). For expository convenience, the problem is briefly described in the following. Fig. 3 shows the topology of the network as well as link properties. We assume $i^d = 1, \forall \; a \in A$, (cf. Eq. (9)), following Ziliaskopoulos (2000). The analysis period is 100 s (T), divided into 10 time intervals (so $A = 10$). The destination node is node F, and the travel demand is $q_{0r}^A = 8, q_{1r}^A = 16, q_{2r}^A = 8$. The original problem also includes an accident on link DE, which reduces its normal capacity (6 per $A$) to zero at $t = 2, 3$ and to 3 (per $A$) at $t = 4, 5$.

A cell-based representation of the network is given in Fig. 4b. In order to avoid infeasibility caused by excessive demands, a dummy origin cell, which has an infinite holding capacity, is added. Whenever the released demands cannot be fully absorbed by the physical network due to capacity restriction, they will be held in dummy origin cells. Similarly, a dummy destination cell is introduced to be a “sink” of all flows. Note that the travel time spent in the dummy destination cell is not included in the objective function. On the contrary, the travel time spent in a dummy origin cell has to be counted; otherwise no traffic will be moved out of that cell. No other changes are made to the original network except that one intra-link intersection is added to each of the links BC and CE. For comparison, Fig. 4a re-plots the CTM-based network employed in Ziliaskopoulos (2000), which also requires adding dummy origin and destination cells. Although the two representations are in fact very similar, their appearance is not. In particular, the CTM-based network emphasizes the cross-cell fluxes (see Fig. 4a). One difference between the two networks is worth noting. In the CTM network, the bottom route (corresponding to A–B–C–E–F) has a free flow travel time of 6 $A$, including 1 $A$ spent in the dummy origin cell. In our network (Fig. 4b),
however, the travel time of the same route is 7 $\Delta$. To make our results comparable to those from Ziliaskopoulos (2000), Cell 10 is removed from link BC for the purpose of verification.

The model (11) based on the above cell representation has 300 solutions variables and 260 constraints (without counting the upper bounds on the solution variables). As a comparison, the model in Ziliaskopoulos (2000) has 200 variables and 550 constraints. The resulting LP problem is solved using "linprog" function in MATLAB 6. To obtain the most reliable performance, we selected the simplex method and turned off the "large-scale" option in MATLAB. Table 1 reports the optimal cell volumes $p_{at}^t / c_0 / c_1$, exit flow $v_{at}^t / c_0 / c_1$, and entry flow $u_{at}^t / c_0 / c_1$. The total system cost is 215, which is exactly same as the value reported in Ziliaskopoulos (2000). Fig. 5 shows that the network status under SO-DTA at $t = 2$ obtained from our model is also comparable to the CTM result (see Fig. 4 in Ziliaskopoulos (2000)) and note that $t = 2$ is actually the third interval since $t$ is

---

4 The CTM representation in Fig. 4a, adopted from Ziliaskopoulos (2000), is slightly inconsistent with the original network given in the same reference. To see this, note that in the original network (Fig. 3), the free flow travel time on path A–B–C–E–F is 6 $\Delta$ whereas in the CTM network (Fig. 4a), the travel time on the same path (without considering the one unit spent in the dummy cell) is 5. Since the original cell-based network (given in Fig. 4b) does not have the same problem, its SO-DTA results would not be comparable to those benchmark solutions reported in Ziliaskopoulos (2000). Therefore, it is necessary to remove cell 10 from link BC for the purpose of verification.
counted from 0 in our study. This equivalence is indeed expected from our analysis, which reveals that both models embed relaxed CTM traffic flow dynamics.

Having verified the proposed model, we now check the necessary condition given in Theorem 1. Specifically, all three routes connecting the node A and the destination F are considered (represented using cell ID given in Fig. 4):

- **k1**: 0–1–3–6–7–8.
- **k2**: 0–1–2–5–6–7–8.
- **k3**: 0–1–2–4–8.

The marginal costs on these paths and their subpaths are computed using the recursive formula (15), (16) for all time intervals. These values are then compared with the multipliers associated with the flow propagation constraints ($\lambda_t^k$) in Table 2. For the convenience of presentation, we ignore the subpath results starting at cells 5–9. Table 2 shows that, at system optimum, the marginal costs of the three paths are identical and equal to the optimal multiplier values in most cases. There are a few cases where different marginal costs are observed on different paths. For example, at time 0 and cell 1, only path $k_3$ has the marginal cost equal to $\lambda_t^k$, which is 4 in this case. Note that if a unit of flow is released on cell 1 at $t = 0$, traveling on

[Table 1]

<table>
<thead>
<tr>
<th>Optimal cell volume ($p_t^a$ + $r_t^a$)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>12</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Optimal queued flows ($p_t^a$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Optimal entry flow ($u_t$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>

*Fig. 5.* Network flow pattern at system optimum when $t = 2$. 
path $k_3$, it will arrive at cell 4 at $t = 2$, before the accident occurs. Consequently, the flow unit can take this shortcut to arrive at the destination in four intervals. Also, this unit of flow should not use the other two paths, because they both are longer. However, the flow released in the same cell at the next time interval will be affected by the accident. The differences observed for cell 2 at $t = 0, 1$ are caused by exactly the same reason.

We finally examine the flow holding back problem. Interestingly, the system optimum solution reported in Table 1 does not contain any unnecessary flow holding back. To reveal how such holding might arise, another accident is introduced to reduce the capacity of link E-F from 12 to 3 during time intervals 4 and 5 (see Fig. 3). The SO cost of the new problem increases, as expected, from 215 to 235. Also expected was the occurrence of the unnecessary flow holding back problem (see Table 3, which reports optimal cell volumes). The table shows that six units of flows always stayed in the first cell. While they could have been moved closer to the destination, the LP solution procedure deems this effort useless since these flows will not make it to the destination anyway due to the newly imposed capacity restriction on link E–F. Nevertheless, this flow holding solution violates the physical rule imposed in the first place, which requires that the capacity should be fully utilized before a queue develops.

To fix the above flow holding problem, the algorithm presented in Section 5 is implemented in MATLAB. The algorithm identified a holding-free solution after just one iteration. The optimal objective function of the first maximization problem (i.e. the maximum exit flows from $t = 0$ to 8) is 190. The optimal cell volumes are reported in Table 4. We have verified that

Table 2
Path marginal costs and corresponding optimal multiplier values $z_k$.

<table>
<thead>
<tr>
<th>Cell</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>$z_k$</td>
<td>$k_1$</td>
<td>$k_2$</td>
<td>$k_3$</td>
<td>$z_k$</td>
</tr>
<tr>
<td>0</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3
System optimum cell volumes with the new accident (unnecessary flow holding back).

<table>
<thead>
<tr>
<th>Time/Cell</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4
System optimum cell volumes with the new accident (no flow holding back).

<table>
<thead>
<tr>
<th>Time/Cell</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
this solution indeed satisfies the OSP constraint (10). Intuitively, one can see that the six units of “stationary” flows in the initial solution have been moved to cells 7 and 8, the closest cells to the destination. This preliminary experiment thus demonstrates the effectiveness of Ho’s algorithm in resolving unnecessary flow holding back. Also, our test agrees with Ho’s observation Ho (1980) that the number of required iterations is much smaller than T.

7. Conclusions

This paper revisits the 30-year old model of Merchant and Nemhauser (1978a) for the system optimum dynamic traffic assignment problem. Despite its popularity, the M-N model has been frequently criticized for the “unrealistic” representation of traffic dynamics by the exit-flow function. While there can be no doubt that many of these criticisms are well motivated and grounded, it has been recognized recently (Carey and McCartney, 2004) that most alleged violations of physical rules may well be a result of inappropriate discretization schemes.

Motivated by the above observations, this paper aims to show that the M-N model can encapsulate realistic traffic flow models such as CTM, as long as the exit-flow function is properly chosen and discretized. To this end, a cell-based variant of the M-N model is first proposed using a simple piecewise linear exit-flow function. Once linearized, this model embeds the point-queue traffic flow model, which can be further extended to incorporate the cross-cell interactions and thereby capturing queue spillover effects. By both analytical and numerical demonstrations, the proposed model was shown to be equivalent to the CTM-based SO-DTA model. It is different, to some extent better, than the existing CTM model because it (1) does not require intersections to be standard merge and diverge, and hence is easier to construct and (2) does not directly solve for cell-to-cell flux and generally has a simpler constraint structure. We proposed a recursive formula for path marginal costs and proved the necessary condition which is a dynamic extension of the Wardrop’s second principle (Wardrop, 1952). While our approach to the marginal cost analysis is along the line of Merchant and Nemhauser (1978b), our model is more complicated because of the involvement of different multipliers. The proposed model also enables a solution to unnecessary flow holding back resulted from the violation of the ordered solution property (OSP). Specifically, we show that Ho’s algorithm for the original M-N model can also be used to tackle the cell-based model. The proposed model, the marginal cost analysis and the resolution of the flow holding back issue are all verified using a numerical example that has been studied in the literature.

We have noted in Section 4 that the marginal cost analysis could be used to design more efficient solution algorithms for the SO-DTA models. If marginal costs can be evaluated for any given path at a given flow pattern, the conventional iterative algorithms for user-equilibrium DTA problems (such as the method of successive average) can be employed to take advantage of the network structure. However, since the proposed recursive formula involves multipliers, one has to solve for multipliers and primal variables simultaneously. Note that some multipliers, such as those associated with capacity constraints, can be easily computed once primal variables are known (e.g. Shen et al., 2007). It would be interesting to further investigate whether all involved multipliers can be evaluated in a similar fashion.

In the numerical experiment, a flow holding back problem was “created” by imposing an extra capacity restriction such that some flows can never arrive at the destination within the assignment horizon. One could argue that this holding problem is artificial in that it arises only because the assignment horizon is not long enough. It is unclear, however, whether or not all unnecessary holdings would be removed when the assignment horizon is set to be long enough. We leave this intriguing question also to the future research.

Acknowledgements

The author would like to thank Dr. Wen Shen for many enlightening discussions. Fig. 1 is a reproduction based on a similar plot in her Ph.D. thesis. The constructive comments of two anonymous referees on an earlier draft are greatly appreciated. This research was partially supported by National Science Foundation (CMMI-0928577).

References


