Volume 45B, issue 6, July 2011
ISSN 0191-2615
ELSEVIER
TRANSPORTATION RESEARCH
AN INTERNATIONAL JOURNAL
Part B: Methodological
Editor-in-Chief: Fred L. Mannering

Transportation Research Part B 45 (2011) 880–902

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

A dual decomposition method for sector capacity constrained traffic flow optimization

D. Sun [a,*], A. Clinet [b], A.M. Bayen [c]

[a] School of Aeronautics and Astronautics, Purdue University, West Lafayette, IN 47907, USA
[b] Direction Générale de l'Aviation Civile (DGAC), Athis-Mons 91200, France
[c] Department of Civil and Environmental Engineering, University of California, Berkeley, CA 94720, USA

ARTICLE INFO

Article history:
Received 30 April 2010
Received in revised form 26 March 2011
Accepted 26 March 2011

Keywords:
Traffic flow management
Dual decomposition
Large scale Linear Program

ABSTRACT

An aggregate air traffic flow model based on a multicommodity network is used for traffic flow management in the National Airspace System. The problem of minimizing the total travel time of flights in the National Airspace System of the United States, subject to sector capacity constraints, is formulated as an Integer Program. The resulting solution achieves optimal delay control. The Integer Program implemented for the scenarios investigated has billions of variables and constraints. It is relaxed to a Linear Program for computational efficiency. A dual decomposition method is applied to solve the large scale Linear Program in a computationally tractable manner. A rounding algorithm is developed to map the Linear Program solution to a physically acceptable result, and is implemented for the entire continental United States. A 2-h traffic flow management problem is solved with the method.

Published by Elsevier Ltd.

1. Introduction

The *National Airspace System* (NAS) in the United States is a large scale, nonlinear dynamic system with a control authority which is organized hierarchically. A single *Air Traffic Control System Command Center* (ATCSCC) in Herndon, VA, supervises the overall traffic flow. Organized by geographical region, the airspace is divided into 22 (20 in the continental US) *Air Route Traffic Control Centers* (ARTCCs, or simply, Centers), controlling the airspace up to 60,000 feet (Nolan, 2003). Each Center is sub-divided into smaller regions, called *Sectors*, with at least one Air Traffic Controller responsible for each sector. The last few decades have witnessed the almost uninterrupted growth of air traffic. Except for a dip immediately after the tragic events of September 11, 2001, air traffic in the United States continues to grow at a steady pace. There are different growth scenarios associated both with the magnitude and the composition of the future air traffic. In particular, the *Terminal Area Forecast* (TAF), prepared every year by the *Federal Aviation Administration* (FAA), projects the growth of traffic in the United States (Anonymous, 2005). Since the main function of Air Traffic Controllers is to maintain safe separation between aircraft while guiding them to their destinations, an imbalance between the growth of air traffic and airspace capacity poses potential safety and efficiency issues to the air traffic control system. Thus, the design of a semi-automated *Air Traffic Control* (ATC) system is promising to help Air Traffic Controllers manage the increasing complexity of traffic flow in the en route airspace.

Current *Traffic Flow Management* (TFM) operations in the United States have two coupled phases. The first phase includes national-level flow management initiatives over a 2–8 h planning horizon. This is administrated by the ATCSCC, interested in overall traffic flow. The goal of national traffic flow management is to account for user-preferred gate-to-gate trajectories by

[*] Corresponding author. Tel.: +1 765 494 5718.
E-mail addresses: dsun@purdue.edu (D. Sun), alexis.clinet@aviation-civile.gouv.fr (A. Clinet), bayen@berkeley.edu (A.M. Bayen).

0191-2615/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.trb.2011.03.004

managing and allocating NAS resources in situations when demand approaches or exceeds supply (Sridhar et al., 2008; Grabbe et al., 2009). In the second phase, Center-level controls are designed for local flows and or refinement of the national-level plan in response to updated weather and traffic information. This Center-level management is performed over a 30 min to 2 h time horizon as a tactical control loop to implement the national-level initiatives and to introduce local flow control initiatives such as miles-in-trail and tactical rerouting (Idris et al., 2006; Vossen et al., 2000; Grabbe et al., 2009). In the current air traffic system, national and tactical flow management heavily relies on human operators, with a limited number of decision support tools such as the *Enhanced Traffic Management System* (ETMS) (Volpe National Transportation Center, 2005) and the *Flight Schedule Monitor* (FSM) (US Department of Transportation and Federal Aviation Administration, 2008). Human operations are based on a combination of intuition and past experience, which can possibly result in inefficient controlling of air traffic flows. Meanwhile, air traffic in the United States is forecasted to double or triple over the next twenty years (Joint Planning and Development Office, 2009). There is thus a need to develop advanced traffic management concepts and techniques that can (partially) automate the control process and therefore maximize the throughput and efficiency of the NAS under limited resources.

In order to evaluate the impact of TFM procedures on overall NAS performance, a number of metrics, such as delay, safety, predictability, access, flexibility, and efficiency, have been proposed to describe the performance of the system (Bradford et al., 2000), and details about the metrics are available in Bolczak et al. (1997), Gosling (1999) and Kopardekar et al. (2009), among which delay has been used extensively to assess the performance of NAS (Callaham et al., 2001; Wang et al., 2003; Bratu and Barnhart, 2005; Chatterji and Sridhar, 2005a; Xu et al., 2005; Laskey et al., 2006; Sridhar and Swei, 2006, 2007; Klein, 2007; Churchill et al., 2007). A number of recent studies have proposed methods to facilitate the development of TFM control strategies. In general, there are two complementary parts for these studies: (i) Models for the air traffic system. (ii) Optimization methods for optimal TFM.

Among air traffic models, methods based on aggregate modeling have been of significant interest to the *Air Traffic Management* (ATM) community in the recent years. Aggregate and reduced order models simplify the analysis and design of numerous complex systems. The dimension and complexity of an aggregate model do not change with respect to the number of flights in general, which is a very useful feature to study benchmark scenarios in which air traffic doubles or triples (Menon et al., 2002; Sridhar et al., 2006; Bayen et al., 2006; Sun and Bayen, 2008). Aggregate models thus contrast with physics-based modeling approaches including the *Center TRACON Automation System* (CTAS) (Erzberger, 1992), the *Future ATM Concepts Evaluation Tool* (FACET) (Bilimoria et al., 2001), and the *Collaborative Routing Coordination Tool* (CRCT) (Rhodes et al., 2001), which are agent-based models. The development of aggregate flow models for air traffic flow management has been the subject of considerable interest since the first model (Menon et al., 2002) appeared in the literature. Previously published aggregate flow models (Sridhar et al., 2006) represent and predict the traffic behavior to a high degree of accuracy and can be tailored to the time-scales and regions of interest. A comparison of the characteristics of the different aggregate flow models can be found in previous studies (Sridhar and Menon, 2005; Sun et al., 2006, 2007). In addition, stability and response characteristics of the aggregate flow models are presented in Chatterji and Sridhar (2005b). The aggregation in flow models generally results in the loss of information about the route structure of individual aircraft. Recent studies have shown that efficient disaggregation methods can be designed to convert flow-based (aggregate) solutions to flight-specific control actions (Sun et al., 2009, 2010).

In parallel, optimization techniques are developed (mostly based on specific or a class of air traffic models) to facilitate the design of TFM strategies. Current popular TFM control schemes have been mainly focused on ground delay and/or rerouting flights to accommodate capacitated elements, en route sectors and airports (Lulli and Odoni, 2007). TFM studies focusing on optimal ground delays have been conducted by many researchers, from both deterministic and probabilistic perspectives (see Odoni, 1987; Terrab and Odoni, 1993; Gilbo, 1993; Vranas et al., 1994; Andreatta et al., 1995; Andreatta and Brunetta, 1998; Navazio and Romanin-Jacur, 1998; Bertsimas and Stock Patterson, 1998; Hoffman and Ball, 2000; Dell'Olmo and Lulli, 2003; Vossen et al., 2003; Ball and Lulli, 2004; Ball et al., 2005; Vossen and Ball, 2005; Vossen and Ball, 2006; Ball et al., 2010), and flight delay propagation has also been studied (Churchill et al., 2010). Odoni formulated the TFM problem using a large number of models and algorithms to detect optimal strategies to assign ground delays to flights (see Odoni, 1987; Bertsimas et al., 2008). Helme was among the first to include en route capacity restrictions in the TFM problem (Helme, 1992), which is intuitive to understand but has weak computational performance as was discussed in Bertsimas et al. (2008). Lindsay et al. formulated a disaggregate deterministic 0–1 integer programming model for deciding ground and airborne holding of individual flights in the presence of both airport and airspace capacity constraints (Lindsay et al., 1993). A deterministic, open-loop integer programming method was formulated to assign departure time and sector occupancy time of each aircraft in the work of Bertsimas and Stock Patterson (1998), but the computational complexity of this model has limited its use to a small number of real-world examples as was shown in Grabbe et al. (2007). To improve the runtime, a method that reduces the number of flights to be optimized was proposed in Rios and Ross (2008); and more recently, a Dantzig–Wolfe decomposition method was implemented for the Bertsimas and Patterson model (Rios and Ross, 2010), which actually motivated several studies (including the work in this paper) using decomposition methods to solve large scale TFM problems. The work in Bertsimas and Stock Patterson (1998) was extended to provide a complete representation of all the phases of each flights including rerouting strategies (Bertsimas et al., 2008). In Sherali et al. (2002), a binary integer programming was proposed for a TFM problem, which considers controller workload, airspace safety, and equity among airlines. Subsequently, the binary IP was extended to incorporate rerouting, in Sherali et al. (2003) and Sherali et al. (2006). Research that considers equity or market-based traffic management using aggregate models has been conducted in Bloem and Sridhar (2008), Waslander et al. (2008a,b). Sridhar et al. proposed an integrated three-step hierarchical method for developing deterministic TFM plans consisting of national-level playbook reroutes, miles-in-trail restrictions, and tactical reroutes to alleviate sector-level congestion (Sridhar et al., 2002). Subsequently, Kopardekar and Green used a deterministic, Center-based system to manually identify congested sectors and compare the trade-offs of implementing altitude capping, local rerouting, departure delays, and time-based metering or miles-in-trail restrictions (Kopardekar and Green, 2005). Wanke and Greenbaum proposed a Monte Carlo-based incremental, probabilistic decision making approach for developing en route traffic management controls in Wanke and Greenbaum (2007). More recently, Grabbe et al. applied a sequential optimization method to manage air traffic flow under uncertainty in airspace capacity and demand (Grabbe et al., 2009).

In the present article, we use the framework of aggregate air traffic models for strategic TFM, in which optimization for TFM can be cast in the form of optimization-based control of a dynamical system evolving on a network. Although the same methodology could be applied for air traffic control in a smaller portion of the airspace (such as a sector, which mainly concerns the separation of aircraft), it is not the main focus of this work. In this proposed study, given flight departures and destinations, we solve the problem of minimizing the total flight time for all the flights in the entire NAS, by assigning optimal en route delay control actions (such as holding patterns, change of speed) to a group of aircraft in specific sectors along their paths, so that the sector capacity will be respected. The formulation is shown to be adaptable for consideration of optimal ground delay control when airports are part of the model. Because of the size of the optimization problem and the accuracy of the model, the problem, formulated as an Integer Program, has billions of variables and constraints. It is relaxed to a large scale Linear Program for computational efficiency. Solving the Linear Program directly is prohibitively complex due to its high dimension. A method based on dual decomposition is applied to solve the problem (Bertsekas, 2002).

The method of dual decomposition has been used since the 1960's with the historical work of Dantzig and Wolfe (1960). A good, modern reference of dual decomposition is Chapter 6 in the book of Bertsekas (2002). The dual decomposition method has been applied in engineering, such as in rate control for communication networks (Kelly et al., 1997), and to networking problems for simultaneous routing and resource allocation (Xiao et al., 2004). Recently, the dual decomposition method was presented in the article Chiang et al. (2007) in an effort towards a systematic understanding of "layering" as "optimization decomposition," where the overall communication network is modeled by a generalized network utility maximization problem, and each layer corresponds to a decomposed subproblem. Alternative decomposition methods were also applied in network utility maximization problems to obtain specific distributed algorithms (Tan et al., 2006; Palomar and Chiang, 2007). The dual decomposition method was also recently used to solve large computationally intractable problems for formation flight with multiple cooperative agents, which resulted in an algorithm that is easily implementable in a decentralized manner (Raffard et al., 2004).

In the present article, it is shown that the dual decomposition method is well suited to the multicommodity network structure of the aggregate traffic flow model, the *Large-capacity Cell Transmission Model*, in short CTM(L), developed in earlier work (Sun and Bayen, 2008). It breaks the large scale Linear Program into a sequence of smaller Linear Program problems (subproblems), which are tractable and can be solved in a reasonably short amount of time. This is helpful for strategic TFM planning/re-planning responsive to different traffic conditions, for example, when the air traffic system is disturbed by unexpected weather perturbations. It consists of an iterative algorithm, in which the subproblems are solved involving their own local variables as well as the variables of subproblems they are coupled with.

The rest of this article is organized as follows. The second section introduces the aggregate air traffic flow model. The third section formulates the optimal traffic flow management problem as a large scale Linear Program. In the fourth section, the Linear Program is solved using a dual decomposition method. Finally, in section five, we present some numerical results.

2. Large-capacity cell transmission model

This section briefly summarizes the modeling framework developed in two earlier articles (Sun et al., 2007; Sun and Bayen, 2008), in which we constructed a traffic flow model used for the present article, the *Large-capacity Cell Transmission Model*, in short CTM(L) (Sun and Bayen, 2008). It is based on a network flow model constructed from historical *Enhanced Traffic Management System* (ETMS) and *Aircraft Situation Display to Industry* (ASDI) air traffic data (Bayen et al., 2006). This model is called Large-capacity Cell Transmission Model in reference to the *Cell Transmission Model* in highway traffic (Daganzo, 1994; Daganzo, 1995). The key element in the CTM(L) most relevant to the *dual decomposition* method is the underlying flow network, which is illustrated in the rest of this section, and uses a multicommodity routing network.

2.1. Definitions

The system modeled in this section is the continental en route US NAS, of the size of 20 ARTCCs, for altitudes 24,000 feet and above. The model can be extended to include airports as is shown in Appendix A.1. The ETMS/ASDI data used in this study, provides the position and altitude of all airborne aircraft in the US, updated every minute. Additional information related to flight plans or other flight parameters, such as speed and heading, are also provided in the data, but were not used to build the present CTM(L).

2.2. Construction of the network

Network models are standard in science and engineering. A network model is composed of fundamental components called vertices or nodes; each node is connected to other nodes by links or edges. The network modeling traffic flow in the NAS in CTM(L) is constructed as follows.

2.2.1. Vertices (nodes)

The vertices in the network are constructed following a paradigm developed in earlier work (Sun and Bayen, 2008). For any two neighboring sectors s_1 and s_2, the vertices at the boundary of s_1 and s_2 are denoted by $v_{\{s_1,s_2\}}$ and $v_{\{s_2,s_1\}}$. Vertex $v_{\{s_1,s_2\}}$ represents the entry point used by flights going from s_1 to s_2, while vertex $v_{\{s_2,s_1\}}$ represents the exit point used by flights going from s_2 to s_1. Because the physical entry or exit points do not affect the modeling (while the lengths of links and connections between links are important, as will be seen in the rest of this section), we assume that the exit point and the entry point on a common boundary between sectors are fixed at the same location (location of a vertex), and that each flight crosses (enters or exits) a sector boundary at the fixed vertex.

2.2.2. Links (edges)

For any sectors s_1, s_2 and s_3, if s_1 and s_2 share a boundary and if s_2 and s_3 are neighbors, two *directed links* are created: one from vertex $v_{\{s_1,s_2\}}$ to vertex $v_{\{s_2,s_3\}}$ and one from vertex $v_{\{s_3,s_2\}}$ to vertex $v_{\{s_2,s_1\}}$. In the rest of this work, the term *link* refers to a directed link (Ahuja et al., 1993). Fig. 1 illustrates the concept of a link.

For each link of the network, the flight times for a full year (October 1st, 2004 to September 30, 2005) extracted ASDI/ETMS data are aggregated into travel time distributions. The mean of this distribution is computed, and its value is chosen to represent the "time length" of the link, i.e. the aggregated travel time along the link. Fig. 2 shows a typical travel time distribution.

2.2.3. Multicommodity network

For a complete network model including the whole continental NAS of the US, a multicommodity flow structure (Cormen et al., 2002) is used in CTM(L). Flights are clustered based on their entry-exit node pairs (origin-destination pairs) in the

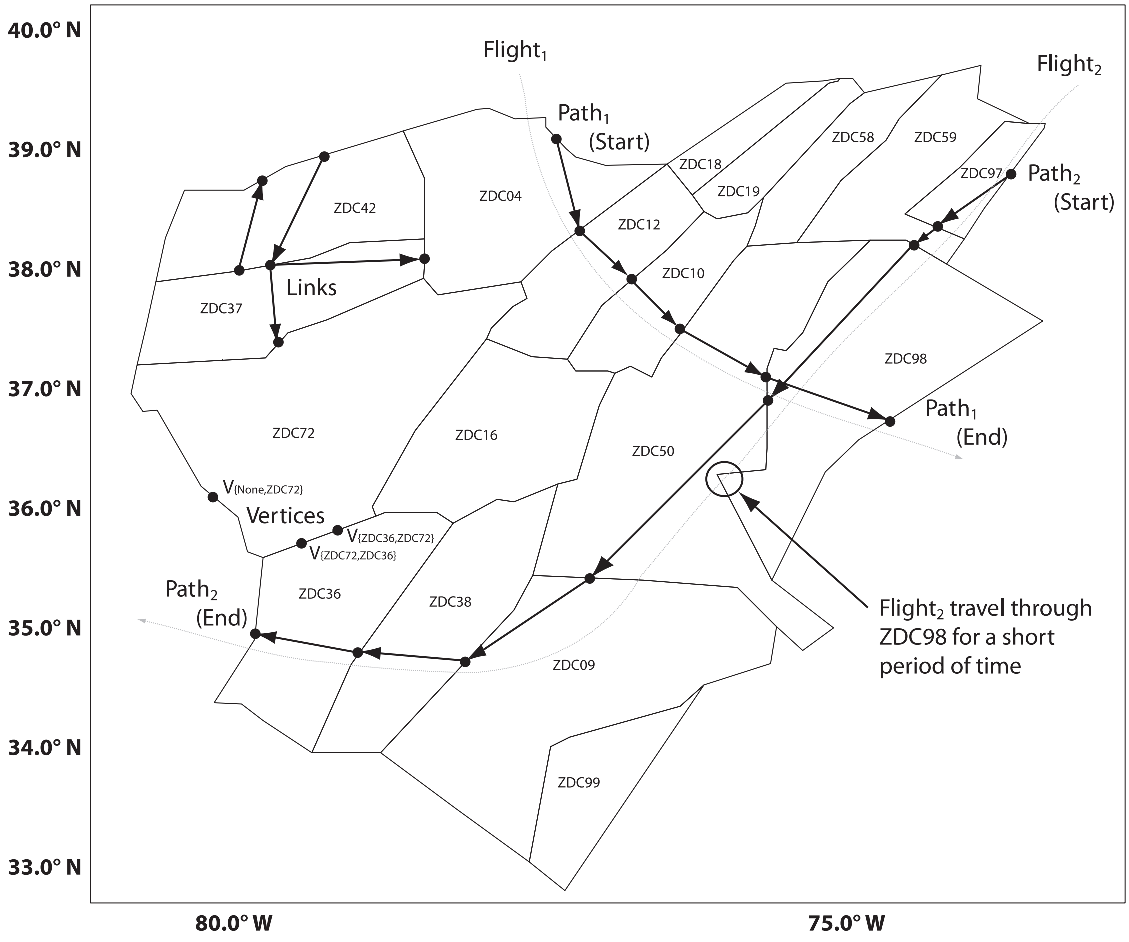

Fig. 1. Examples of vertices, links, trajectories and paths for a subset of the NAS.

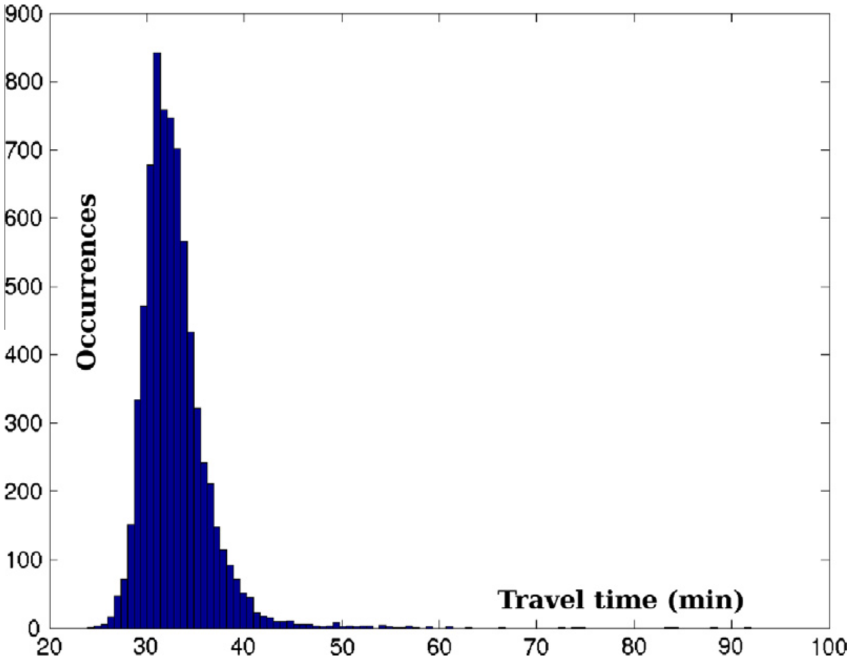

Fig. 2. Distribution of travel time on one link (ZLC45–ZOA33–ZOA34). One full year of aggregated data.

network. Each pair corresponds to a *path* consisting of links between these nodes. If two or more paths have one link in common, this link will be duplicated, using a multicommodity flow structure.

2.3. Dynamics

The CTM(L) model, based on the network constructed above, is developed inspired by the *Lighthill–Whitham–Richards* theory (Lighthill and Whitham, 1956; Richards, 1956), and the Daganzo's *Cell Transmission Model* (CTM) (Daganzo, 1994; Daganzo, 1995), commonly used in highway traffic modeling. The model is reduced to a linear time invariant dynamical system, in which the state is a vector of aggregate aircraft counts. The behavior of air traffic flow on a single link indexed by i can be modeled by a deterministic linear dynamical system with a unit time delay, defined as follows (Sun and Bayen, 2008):

$$x_i(t+1) = A_i x_i(t) + B_1^i u_i(t) + B_2^i f_i(t), \tag{1}$$

$$y(t) = \widetilde{C}_i x_i(t), \tag{2}$$

where $x_i(t) = [x_i^1(t); \ldots ; x_i^{m_i}(t)]$ is the state vector whose elements represent the corresponding aircraft counts in each cell of link i at time step t, and m_i is the number of cells in link i. The forcing input, $f_i(t)$, is a scalar that denotes the entry count onto link i during a unit time interval from t to $t + 1$, and the control input, $u_i(t)$, is an $m_i \times 1$ vector, representing airborne delay control, which can take several forms such as speed change, *vector for spacing* (VFS), *holding pattern* (HP) (Nolan, 2003). Particularly, the update time interval (1 min in this paper) matches the minimum delay control time unit in practice in most cases (Nolan, 2003; Bilimoria et al., 2001). For example, if $u_i(t) = 2$, it means at time t there are two flights subject to airborne delay in cell i with each flight delayed by 1 min. If a longer delay (say, 3 min during the time t_1 to $t_1 + 2$) for two flights in cell i should be actioned, the control variable $u_i(t)$ takes the value $u_i(t) = 2$ for $t = t_1$, $t_1 + 1$ and $t_1 + 2$.

The output, $y(t)$, is a scalar which represents the aircraft count in a user-specified set of cells (indicated by vector $\widetilde{C}_i$) at time step t. The nonzero elements of the $m_i \times 1$ vector $\widetilde{C}_i$ correspond to the cells in the user-specified set, and are equal to one. A_i is an $m_i \times\ m_i$ nilpotent matrix with 1's on its super-diagonal. $B_2^i = [0; \ldots ; 0; 1]$ is the forcing vector with m_i elements, and B_1^i is the $m_i \times m_i$ holding pattern matrix, in which all nonzero elements are 1 on the diagonal and -1 on the super-diagonal.

Based on the link level model, it is straightforward to build a sector level model using the same technique. Suppose that there are n links in a sector, then the dynamics for the sector level model can be described as:

$$x(t+1) = Ax(t) + B_1 u(t) + B_2 f(t), \tag{3}$$

$$y(t) = \widetilde{C} x(t), \tag{4}$$

where $x(t) = [x_1(t); \ldots ; x_n(t)]$ denotes the state, and $f(t) = [f_1(t); \ldots ; f_n(t)]$ is the forcing input vector (the entry count onto the sector). The control input vector is denoted as $u\ (t) = [u_1(t); \ldots ; u_n(t)]$. The scalar $y(t)$ represents the aircraft count in a user-specified set of cells at time step t, and $\widetilde{C}$ corresponds to the cells in the user-specified set, The matrices A, B_1, and B_2 are block diagonal, with block elements associated with each link in the sector. For example, $A = \text{diag}(A_1, \ldots, A_n)$ with A_i's defined by Eq. (1). In the above model, the matrices A, B_1 and B_2 are sparse and structured (e.g., A is nilpotent), which can be exploited to develop efficient algorithms for optimization using this model.

When building a NAS-wide model (at the ARTCC level), flights are first clustered based on their origin-destination (source-sink) pairs in the network. Each pair corresponds to a *path* consisting of links between these nodes. If two or more paths have one link in common, this link will be duplicated. Therefore, the NAS-wide model can also be cast in the same framework of (3) and (4) and the corresponding $x(t)$ includes all cells of the complete network. The forcing input, $f(t)$, is now the entry count into the NAS. Further details about the CTM(L) are described in Sun and Bayen (2008).

2.4. Remarks

- The CTM(L) can be extended to include airports as part of the model. The dual decomposition method in the next sections can be extended and applied for NAS-wide TFM including airports. See Appendix A.1 for details.
- Advantages of CTM(L): The CTM(L) is computationally efficient, and its computational complexity does not depend on the number of aircraft, but only on the size of the network under consideration. Therefore it is suitable to study the air traffic control system in benchmark scenarios for which traffic could double or triple. These scenarios have been of high interest in the air traffic control community. Its control theoretic structure enables the use of standard methods for analysis and optimization. For example, the stability, controllability, observability, and response characteristics of CTM(L) can be directly addressed using control theory (Robelin et al., 2006; Sun and Bayen, 2008). The optimization method using Linear Program/Integer Program which is proposed in the next sections, is based on the fact that the CTM(L) is a linear system model.
- Limitations of CTM(L): Although the CTM(L) is capable of routing flights, i.e., given an origin-destination pair, there could be different routes based on the underlying network representation of the NAS (see Sun and Bayen, 2008, Section 3.2). However, the dual decomposition method proposed in the following sections is limited to delay controls, without rerouting. As a type of aggregate modeling, information about the route structure of individual aircraft is not modeled or preserved. These limitations are addressed in the authors' other work (Sun et al., 2009, 2010). However, the aggregation feature of CTM(L) is useful to support FAA or the operator of the airspace, particularly for resource allocation in the en route airspace for TFM purposes. The disaggregation method will be devised when translating flow-based controls to flight-based controls becomes necessary (Sun et al., 2009, 2010). Re-routing is under investigation and is a key part of future work, which can be done by taking into account a limited number of possible routes for an origin-destination pair, which is a common practice in air traffic control.

3. Problem formulation

3.1. Notations, nomenclature

The following notations are used to formulate the optimization problem presented in the rest of the article.

- $S = \{s_1, s_2, \ldots, s_{|S|}\}$: set of sectors in the NAS. $|S|$ denotes total number of sectors in the NAS.
- $E = \{e_1, e_2, \ldots, e_{|E|}\}$: set of links (Section 2.2). Link $e_m = (v_i, v_j)$, or simply $e_m = (i, j)$, corresponds to an ordered pair of vertices.
- $K = \{k_{o_1,d_1}, k_{o_2,d_2}, \ldots, k_{o_{|K|},d_{|K|}}\}$, or simply $K = \{k_1, k_2, \ldots, k_{|K|}\}$: set of origin-destination (OD) pairs: origin (source) o_k, destination (sink) d_k. $|K|$ denotes total number of OD pairs. In this article, OD pairs can also be understood as "paths," because given an OD pair, a path is uniquely defined: no multiple paths exist between a single OD pair.[1]
- T: time horizon of the optimization.
- $Q_s \subset E$: set of cells in sector $s \in S$. $(j, k) \in Q_s$ means the jth cell on path k is in Q_s (therefore in sector s).
- h_t^k: scheduled input (departure) flights into path k at time t. The vector h_t is used to denote the aggregation of scheduled departures to the entire NAS at time t.
- $x_t^{j,i}$: state of cells in the dynamical system. $x_t^{j,i}$ represents the number of aircraft in the jth cell on path i (namely cell (i, j); defined in Section 2.2) at time t. The vector x_t^k is used to denote the aggregation of states on path k at time t: $x_t^k = [x(k,1,t); x(k,2,t); \ldots; x(k,n(k),t)]$, where $n(k)$ is the total number of cells on path k. The vector x^k is the aggregation of states on path k: $x^k = [x_1^k; x_2^k; \ldots; x_T^k]$. The vector x_t is the aggregation of states at time t: $x_t = [x_t^1; x_t^2; \ldots; x_t^k; \ldots; x_t^{|K|}]$. The vector x is the vector of all the states: $x = [x(1); x(2); \ldots; x(T)]$.
- $x_t^{r_i}$: state of last cell (the r_ith cell) on path i at time t, where r_i is the total number of cells on path i. The vector $x_t^r = [x_t^{r_1}; x_t^{r_2}; \ldots; x_t^{r_k}; \ldots; x_t^{r_{|K|}}]$ is the aggregation of states for last cells on each path.
- x_p^k: sum of all airborne flights on path k at time $t = 0$: $x_p^k = \sum_{j=1}^{r_k} x_0^{k,j}$. The vector $x_p = [x_p^1; x_p^2; \ldots; x_p^k; \ldots; x_p^{|K|}]$ is the aggregation of the sums.
- $u_t^{j,i}$: delay control in cell (i, j) at time t, representing number of delay controlled aircraft in the jth cell on path i at time t. The vector u_t^k is used to denote the aggregation of controls on path k at time t: $u_t^k = [u(k,1,t); u(k,2,t); \ldots; u(k,n(k),t)]$,

[1] This feature can easily adapt to cases of multiple paths between an OD pair. It is omitted in this article for simplicity of the description.

where $n(k)$ is total number of cells on path k. The vector u^k is the aggregation of the controls on path k: $u^k = [u^k_1; u^k_2; \ldots; u^k_T]$. The vector u_t is the aggregation of controls at time t: $u_t = [u^1_t; u^2_t; \ldots; u^k_t; \ldots; u^{|K|}_t]$. u represents the vector of all the controls: $u = [u(1); u(2); \ldots; u(T)]$.

- $c(i, j)$, $i = 1, \ldots, |K|$, $j = 1, \ldots, n_i$ (n_i is the number of cells on path i), means the cost associated with flying through cell (i, j), which is the travel time of a flight through cell (i, j). $c(i, j) = 1$ represents 1 min travel time in the present CTM(L). Let c^k represent an aggregation of costs on path k: $c^k = [c(k,1); c(k,2); \ldots; c(k, n(k))]$, and let $c = [c^1; c^2; \ldots; c^{|K|}]$.
- $C_s(t)$, $s = 1, \ldots, |S|$: sector capacity for sector s at time t. The sector capacities are time dependent because the usage of airspace is dynamic: capacity can change due to weather or operations. Denote $C(t) = [C_1(t); C_2(t); \ldots; C_{|S|}(t)]$, as the aggregation of capacity constraints at time t, and $C = [C(1); C(2); \ldots; C(T)]$ as the aggregation of all sector capacities at all times.
- Slack variables $Z^k_s(t)$, $s = 1, \ldots, |S|$, $k = 1, \ldots, |K|$, represent the number of aircraft in sector s on path k at time t: $Z^k_s(t) = \sum_{(j,k) \in Q_s} x(k, j, t)$, where (j, k) is the jth cell on path k and Q_s is the set of links in sector s. Let $Z^k(t)$ denote the aggregation of slack variables $Z^k_s(t)$ on path k at time t: $Z^k(t) = [Z^k_1(t); Z^k_2(t); \ldots; Z^k_{|S|}(t)]$, and $Z^k = [Z^k(1); Z^k(2); \ldots; Z^k(T)]$ are slack variables associated with path k. Let $Z(t)$ denote the aggregation of all slack variables at time t: $Z(t) = [Z^1(t); Z^2(t); \ldots; Z^{|K|}(t)]$. Let $Z = [Z(1); Z(2); \ldots; Z(T)]$ be the aggregation of all slack variables.
- $\mathbb{T}_0 = \{0, \ldots, T-1\}$: set of time indices from 0 to $T - 1$.
- $\mathbb{T} = \{0, \ldots, T\}$: set of time indices from 0 to T.
- $\mathbb{S} = \{1, \ldots, |S|\}$: set of sector indices.
- $\mathbb{K} = \{1, \ldots, |K|\}$: set of path indices.
- $\mathbb{Z}_+$: the non-negative integer set.

3.2. Formulation of en route airspace problem

The problem of minimizing the total travel time of the flights in the NAS can be formulated as follows:

$$\min_{x,u} \quad \sum_{t=0}^{T} c^T x_t, \tag{5}$$

$$\text{s.t.} \quad x_0 = B_2 f_0, \tag{6}$$

$$x_{t+1} = A x_t + B_1 u_t + B_2 f_t, \qquad t \in \mathbb{T}_0, \tag{7}$$

$$\sum_{t'=0}^{t} f_{t'} \leqslant \sum_{t'=0}^{t} h_{t'}, \qquad t \in \mathbb{T}_0, \tag{8}$$

$$\sum_{t'=0}^{T} f_{t'} = \sum_{t'=0}^{T} h_{t'}, \tag{9}$$

$$\sum_{t'=0}^{T} x^r_{t'} = \sum_{t'=0}^{T} h_{t'} + x_p, \tag{10}$$

$$\sum_{(i,j) \in Q_s} x^{i,j}_t \leqslant C_s(t), \qquad s \in \mathbb{S}, \quad t \in \mathbb{T}, \tag{11}$$

$$u \leqslant x \tag{12}$$

$$x \subset \mathbb{Z}_+ \tag{13}$$

$$u \subset \mathbb{Z}_+. \tag{14}$$

The objective function (5) encodes the minimization of the total travel time for all the flights in the NAS for the time horizon of interest. Eq. (6) represents the initial condition, i.e., the airborne flights at the beginning of optimization. Eq. (7) encodes the dynamics of the system. Eq. (8) ensures that accumulated departures (left-hand side of the equation) cannot exceed the amount of scheduled departures (right-hand side of the equation). Eq. (9) enforces all scheduled flights will departure, while Eq. (10) enforces all flights will land at their destinations by end of the planning time horizon. Constraints (8)–(10) are important in TFM optimization problems when the departures are variables to be optimized (Sun et al., 2010). Eq. (11) enforces the capacity constraint for every sector, meaning that the number of aircraft in the sector cannot exceed the sector capacity. Equation (12) is the control constraint for every cell: the number of delay controlled aircraft cannot exceed the total number of aircraft in the cell. Eqs. (13) and (14) represent the non-negativity integer constraint on the states and controls, respectively.

3.3. Remark

Using the framework above, different objective functions can be used for other optimization purposes. In particular, as long as the objective function is convex, one can find efficient algorithms to solve the optimization problem (Boyd and Vandenberghe, 2004; Bertsekas, 2002). Moreover, when the terms in the objective function are separable path by path, the dual decomposition method described in the rest of this article can be applied following the same algorithm developed in Section 4. This is the main contribution of the article, which makes this large scale optimization program computationally tractable.

4. Dual decomposition

A realistic and accurate NAS model approximately includes 100,000 paths to accurately represent the high altitude flow structure of air traffic. Every path usually has hundreds of cells. In order to perform 2-h TFM ($t = 1,\ldots,120$), the problem consists of about five billion states and controls ($x(i, j, t)$ and $u(i, j, t)$); the number of constraints is of the same order (billions). The formulation in Section 3.2 is an *Integer Program* (IP), which is computationally challenging to solve efficiently. To make the problem tractable, we relax the last two constraints (13) and (14) to $u \geqslant 0$, and as a consequence, $x \geqslant u \geqslant 0$, by constraint (12). The formulation now becomes a *Linear Program* (LP):

$$\min_{x,u} \quad \sum_{t=0}^{T} c^T x_t, \tag{15}$$

$$\text{s.t.} \quad x_0 = B_2 f_0, \tag{16}$$

$$x_{t+1} = Ax_t + B_1 u_t + B_2 f_t, \qquad t \in \mathbb{T}_0, \tag{17}$$

$$\sum_{t'=0}^{t} f_{t'} \leqslant \sum_{t'=0}^{t} h_{t'}, \qquad t \in \mathbb{T}_0, \tag{18}$$

$$\sum_{t'=0}^{T} f_{t'} = \sum_{t'=0}^{T} h_{t'}, \tag{19}$$

$$\sum_{t'=0}^{T} x_{t'}^{r} = \sum_{t'=0}^{T} h_{t'} + x_p, \tag{20}$$

$$\sum_{(i,j)\in Q_s} x_t^{i,j} \leqslant C_s(t), \qquad s \in \mathbb{S}, \quad t \in \mathbb{T}, \tag{21}$$

$$0 \leqslant u_t \leqslant x_t, \qquad t \in \mathbb{T}. \tag{22}$$

However, LP relaxation does not change the size of the problem: we still have the same number of variables and constraints as in the IP formulation. Now, we apply the dual decomposition method (Bertsekas, 2002) to solve the large scale LP.

Step 1 Decompose the terms path by path. The objective function can be rewritten as a summation of the total travel time of flights along each path, where the path index is denoted by k. Each constraint can also be written path by path, which is also indexed by k for the kth path.

$$\min_{x,u} \quad \sum_{k=1}^{|K|} \left(\sum_{t=0}^{T} {c^k}^T x_t^k \right),$$

$$\text{s.t.} \quad x_0^k = B_2^k f_0^k, \qquad k \in \mathbb{K},$$

$$x_{t+1}^k = A^k x_t^k + B_1^k u_t^k + B_2^k f_t^k, \qquad t \in \mathbb{T}_0, \quad k \in \mathbb{K},$$

$$\sum_{t'=0}^{t} f_{t'}^k \leqslant \sum_{t'=0}^{t} h_{t'}^k, \qquad t \in \mathbb{T}_0,$$

$$\sum_{t'=0}^{T} f_{t'}^k = \sum_{t'=0}^{T} h_{t'}^k,$$

$$um_{t'=0}^{T} x_{t'}^{r_k} = \sum_{t'=0}^{T} h_{t'}^k + x_p^k,$$

$$0 \leqslant u_t^k \leqslant x_t^k, \qquad t \in \mathbb{T}, \quad k \in \mathbb{K},$$

$$\sum_{k=1}^{|K|} \sum_{(i,k)\in Q_s} x_t^{i,k} \leqslant C_s(t), \qquad s \in \mathbb{S}, \quad t \in \mathbb{T}.$$

Step 2 Introduce slack variables $Z_s^k(t)$, $Z(t)$ and Z, as defined in Section 3.1.

$$\min_{x,u,Z} \quad \sum_{k=1}^{|K|} \left(\sum_{t=0}^{T} {c^k}^T x_t^k \right),$$

$$\text{s.t.} \quad x_0^k = B_2^k f_0^k, \qquad k \in \mathbb{K},$$

$$x_{t+1}^k = A^k x_t^k + B_1^k u_t^k + B_2^k f_t^k, \qquad t \in \mathbb{T}_0, \quad k \in \mathbb{K},$$

$$\sum_{t'=0}^{t} f_{t'}^k \leqslant \sum_{t'=0}^{t} h_{t'}^k, \qquad t \in \mathbb{T}_0$$

$$\sum_{t'=0}^{T} f_{t'}^k = \sum_{t'=0}^{T} h_{t'}^k,$$

$$\begin{aligned}
&\sum_{t'=0}^{T} x_{t'}^{r_k} = \sum_{t'=0}^{T} h_{t'}^k + x_p^k,\\
&0 \leqslant u_t^k \leqslant x_t^k, \qquad t \in \mathbb{T}, \quad k \in \mathbb{K}\\
&\sum_{(i,k)\in Q_s} x_t^{i,k} = Z_s^k(t), \qquad s \in \mathbb{S}, \quad k \in \mathbb{K}, \quad t \in \mathbb{T},\\
&\sum_{k=1}^{|K|} Z_s^k(t) \leqslant C_s(t), \qquad s \in \mathbb{S}, \quad t \in \mathbb{T}
\end{aligned}$$

Step 3 By forming the partial Lagrangian for the last constraints and switching the min and max operators, we can obtain the dual problem:

$$d^* := \max\nolimits_{\lambda \geqslant 0} \; \min\nolimits_{x,u,Z} \sum_{k=1}^{|K|} \left(\sum_{t=0}^{T} {c^k}^T x_t^k \right) + \sum_{t=0}^{T} \sum_{s=1}^{|S|} \lambda_s(t) \left(\sum_{k=1}^{|K|} Z_s^k(t) - C_s(t) \right) \tag{23}$$

$$\begin{aligned}
\text{s.t.} \quad & x_0^k = B_2^k f_0^k, \qquad k \in \mathbb{K},\\
& x_{t+1}^k = A^k x_t^k + B_1^k u_t^k + B_2^k f_t^k, \qquad t \in \mathbb{T}_0, \quad k \in \mathbb{K}\\
& \sum_{t'=0}^{t} f_{t'}^k \leqslant \sum_{t'=0}^{t} h_{t'}^k, \qquad t \in \mathbb{T}_0\\
& \sum_{t'=0}^{T} f_{t'}^k = \sum_{t'=0}^{T} h_{t'}^k,\\
& \sum_{t'=0}^{T} x_{t'}^{r_k} = \sum_{t'=0}^{T} h_{t'}^k + x_p^k,\\
& 0 \leqslant u_t^k \leqslant x_t^k, \qquad t \in \mathbb{T}, \quad k \in \mathbb{K}\\
& \sum_{(i,k)\in Q_s} x_t^{i,k} = Z_s^k(t), \qquad s \in \mathbb{S}, \quad k \in \mathbb{K}, \quad t \in \mathbb{T},
\end{aligned}$$

where d^* denotes the optimal value of the dual problem. Since the problem has linear constraints, the optimal values of the dual problem (23) and the primal problem are equal (Bazaraa et al., 2006). This allows us to solve the primal via the dual.

Step 4 Re-arrange the terms in the objective function of the dual problem (23) to group the terms path by path:

$$\max\nolimits_{\lambda \geqslant 0} \; \min\nolimits_{x,u,Z} \sum_{k=1}^{|K|} \left(\sum_{t=0}^{T} {c^k}^T x_t^k \right) + \sum_{t=0}^{T} \sum_{s=1}^{|S|} \lambda_s(t) \left(\sum_{k=1}^{|K|} Z_s^k(t) - C_s(t) \right) = \max\nolimits_{\lambda \geqslant 0} \left\{ - \sum_{t=0}^{T} \sum_{s=1}^{|S|} \lambda_s(t) C_s(t) + \sum_{k=1}^{|K|} {d^k}^*(\lambda) \right\},$$

where

$${d^k}^*(\lambda) = \min\nolimits_{x,u,Z^k} \sum_{t=0}^{T} {c^k}^T x_t^k + \sum_{t=0}^{T} \sum_{s=1}^{|S|} \lambda_s(t) Z_s^k(t),$$

which is actually the subproblem for path k.

Step 5 Iterations.

- Subproblems ${d^k}^*(\lambda), k \in \mathbb{K}$

$$\min\nolimits_{x,u,Z^k} \quad \sum_{t=0}^{T} {c^k}^T x_t^k + \sum_{t=0}^{T} \sum_{s=1}^{|S|} \lambda_s(t) Z_s^k(t) \tag{24}$$

$$\begin{aligned}
\text{s.t.} \quad & x_0^k = B_2^k f_0^k\\
& x_{t+1}^k = A^k x_t^k + B_1^k u_t^k + B_2^k f_t^k, \qquad t \in \mathbb{T}_0\\
& \sum_{t'=0}^{t} f_{t'}^k \leqslant \sum_{t'=0}^{t} h_{t'}^k, \qquad t \in \mathbb{T}_0\\
& \sum_{t'=0}^{T} f_{t'}^k = \sum_{t'=0}^{T} h_{t'}^k,\\
& \sum_{t'=0}^{T} x_{t'}^{r_k} = \sum_{t'=0}^{T} h_{t'}^k + x_p^k,\\
& 0 \leqslant u_t^k \leqslant x_t^k, \qquad t \in \mathbb{T}\\
& \sum_{(i,k)\in Q_s} x_t^{i,k} = Z_s^k(t), \qquad s \in \mathbb{S}, t \in \mathbb{T}.
\end{aligned}$$

There are $|K|$ (number of paths) subproblems.

Construction of feasible primal variables. It is noticed in this step that optimal variables x (denoted as $x^{\text{opt}}_{\text{dual}}$) derived in the subproblems (24) via the dual approach still violate the dualized constraints (21) during many iterations. It is necessary to construct feasible primal variables, namely x and u (see Bazaraa et al., 2006, Chapter 6). A procedure is designed to solve this problem as follows.

To construct feasible primal variables, when constraint (21) is violated by $x^{\text{opt}}_{\text{dual}}$ (in this case we call that sector s is saturated), the first step is to have the overall sector capacity $C_s(t)$ in (21) divided and allocated to all the paths that pass sector s. The allocated capacity, denoted as $C^i_s(t)$ for path i at time t, is proportional to the amount of flights on the path based on $x^{\text{opt}}_{\text{dual}}$. For example, if there are N paths (indexed by i_1 through i_N) passing sector s and suppose the number of flights in this sector for path i_j is $y_{i_j}(t)$ at time t, where

$$y_{i_j}(t) = \sum_{(i_j,n)\in Q_s} x^{\text{opt}}_{\text{dual}}(t, i_j, n),$$

where $x^{\text{opt}}_{\text{dual}}(t, i_j, n)$ is the solution for the number of flights in the nth cell on path i_j at time t, derived in the subproblem (24) via the dual approach, then the portion of the capacity allocated to path i_j is defined by

$$C^{i_j}_s(t) = \frac{y_{i_j}(t)}{\sum_{(i,n)\in Q_s} x^{\text{opt}}_{\text{dual}}(t,i,n)} \cdot C_s(t). \tag{25}$$

where the denominator $\sum_{(i,n)\in Q_s} x^{\text{opt}}_{\text{dual}}(t,i,n)$ is actually the total number of flights in sector s at t, optimized by (24). Clearly, for these saturated sectors, we have

$$\sum_{k=1}^{|K|} C^k_s(t) = C_s(t). \tag{26}$$

For the sectors whose capacities are not violated by $x^{\text{opt}}_{\text{dual}}$, we define nominal allocated capacities by

$$C^{i_j}_s(t) = y_{i_j}(t). \tag{27}$$

Obviously, in these unsaturated sectors,

$$\sum_{k=1}^{|K|} C^k_s(t) \leqslant C_s(t). \tag{28}$$

Summarizing (26) and (28), we have

$$\sum_{k=1}^{|K|} C^k_s(t) \leqslant C_s(t). \tag{29}$$

The allocated capacities $C^k_s(t)$ will be used in the following step to construct feasible primal variables, in which we implement a quadratic programming-based scheme. Denote the elements associated with the kth path at time t in $x^{\text{opt}}_{\text{dual}}$ as

$$x^{\text{opt}}_{\text{dual}}(t,k) := x^{\text{opt}}_{\text{dual}}(t,k,\cdot),$$

we solve the following quadratic programming:

$$\begin{aligned}
\min_{x,u} \quad & \sum_{t=0}^{T} \left\| x^k_t - x^{\text{opt}}_{\text{dual}}(t,k) \right\|^2_2 \qquad (30)\\
\text{s.t.} \quad & x^k_0 = B^k_2 f^k_0 \\
& x^k_{t+1} = A^k x^k_t + B^k_1 u^k_t + B^k_2 f^k_t, \qquad t \in \mathbb{T}_0 \\
& \sum_{t'=0}^{t} f^k_{t'} \leqslant \sum_{t'=0}^{t} h^k_{t'}, \qquad t \in \mathbb{T}_0 \\
& \sum_{t'=0}^{T} f^k_{t'} = \sum_{t'=0}^{T} h^k_{t'}, \\
& \sum_{t'=0}^{T} x^{r_k}_{t'} = \sum_{t'=0}^{T} h^k_{t'} + x^k_p, \\
& 0 \leqslant u^k_t \leqslant x^k_t, \qquad t \in \mathbb{T} \\
& \sum_{(i,k)\in Q_s} x^{i,k}_t \leqslant C^k_s(t), \qquad s \in \mathbb{S}, \quad t \in \mathbb{T}.
\end{aligned}$$

The solution to the above problem (30) has the following properties:

1. The solution x will be close to $x_{\text{dual}}^{\text{opt}}$ (because of the objective function (30); we do not want x far off the solution to the dualized problem (24), and it always respects the constraints (21) in the primal problem because

$$\sum_{(i,j)\in Q_s} x_t^{i,j} = \sum_{k=1}^{|K|} \underbrace{\sum\nolimits_{(i,k)\in Q_s} x_t^{i,k}}_{\leqslant C_s^k(t)} \leqslant \underbrace{\sum\nolimits_{k=1}^{|K|} C_s^k(t) \leqslant C_s(t)}_{\text{by }(29)}.$$

2. In most cases (which is true for most practical TFM problems), there is a feasible solution to problem (30): we can always hold the flights on the ground (by using small forcing inputs f_t^k) to make sure that the last constraint in (30) is not violated. In the extreme case when the allocated capacity $C_s^k(t)$ is too low which causes infeasibility to problem (30), we can increase $C_s^k(t)$ and decrease $C_s^{k'}(t)$ (or multiple allocated capacities) associated with path k' (or multiple other paths) in sector s, but we still make sure that (29) is satisfied. This is actually equivalent to allocating the capacities $C_s^k(t)$ non-proportional to the amount of flights on path k in sector s based on $x_{\text{dual}}^{\text{opt}}$. If none of such non-proportionally allocated capacities results in a feasible solution to (30), it is because the original TFM problem (5)–(14) does not have a feasible solution, which has not happened in practice.

Fig. 3 shows the percentage of violations of (21). It can be seen that the number of violations converges to zero in around 30 iterations.

- Master problem

$$d^*(\lambda) = \max_{\lambda\geqslant 0}\left\{-\sum_{t=0}^{T}\sum_{s=1}^{|S|}\lambda_s(t)C_s(t) + \sum_{k=1}^{|K|} d^{k^*}(\lambda)\right\}, \tag{31}$$

To solve the dual problem (23), or (31), we need to compute the subgradient of $d^*(\lambda)$. In the subgradient method used in this article, we start with initial $\lambda = \lambda_0 > 0$. At each iteration step $i = 1, 2, 3, \ldots$, we compute a subgradient of the dual function:

$$g(t) = \sum_{k=1}^{|K|} Z^k(t) - C(t), \quad t = 0, \ldots, T.$$

Then we update the dual variable (Lagrange multiplier, in this formulation) by

$$\lambda(t) := (\lambda(t) + \alpha_i g(t))_+, \quad t = 0, \ldots, T,$$

where $(\cdot)_+$ denotes the non-negative part of a vector (i.e., projection onto the non-negative orthant), and α_i is the subgradient step size rule, which is any nonsummable positive sequence that converges to zero (see Boyd and Vandenberghe, 2004; Bazaraa et al., 2006, Chapter 8):

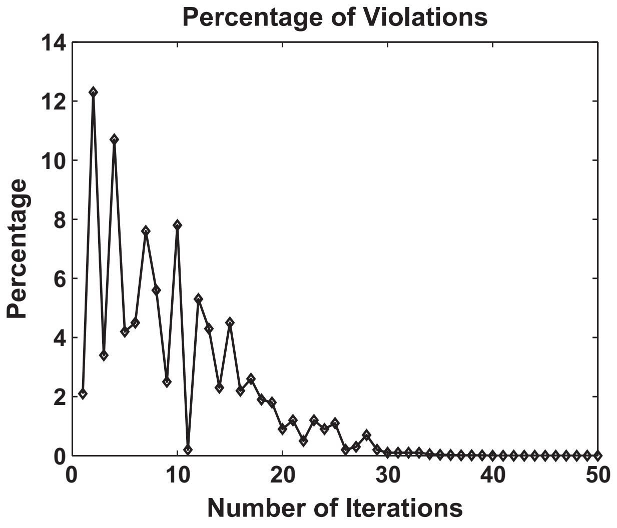

Fig. 3. Percentage of violations of constraints (21) during the iteration of the dual decomposition method.

$$\alpha_i \to 0, \quad \sum_{i=1}^{\infty} \alpha_i = \infty.$$

In this study, we tried several diminishing step size rules: $\alpha_i = 1/(i+1), \alpha_i = 0.5/(i+1), \alpha_i = 0.2/(i+1), \alpha_i = 0.1/\sqrt{i+1}$ and $\alpha_i = 0.02/\sqrt{i+1}$, and we notice that $\alpha_i = 0.02/\sqrt{i+1}$ yielded a better convergence rate, which was therefore chosen as the step size rule throughout this paper. It should be noted that the step sizes we tried are all based on the divergent series; Chapter 2 of Bazaraa et al. (2006) provides several simple yet effective deflected subgradient schemes that will be investigated for future implementation. A proof of convergence of the above algorithm can be found in Chapter 2 of Shor et al. (1985). Numerous methods to accelerate the convergence can be found in the literature (Shor et al., 1985; Bertsekas, 2002).

The dual decomposition method, with the subgradient method for the master problem outlined above, is summarized in Algorithm 1 below.

Algorithm 1. Dual decomposition algorithm.

```
Inputs:
  Initial state $x_0$.
  Time horizon $T$.
  Inputs $f_t, t \in \mathbb{T}$.
  Required sector capacity constraints $C_s(t), s \in \mathbb{S}, t \in \mathbb{T}$.
  Initial $\lambda := \lambda_0 \geqslant 0$.
Start: Iteration number $i = 0$.
repeat
  $i := i + 1$.
  Solve the subproblems (24) for $d^{k*}(\lambda)$, obtain $x_t^k, u_t^k, Z^k(t), k \in \mathbb{K}, t \in \mathbb{T}$.
  Construct feasible primal variables
  Master algorithm subgradients
  $g(t) = \sum_{k=1}^{|K|} Z^k(t) - C(t), t \in \mathbb{T}$.
  Master algorithm update
  $\lambda(t) := (\lambda(t) + \alpha_i g(t))_+, t \in \mathbb{T}$.
until $d^*$ converges or $i$ = max iterations.
Output: $x_t^k, u_t^k, t \in \mathbb{T}, k \in \mathbb{K}$.
```

4.1. A discussion on the integrality of the solution

As observed from our experiments, the optimal solutions to the relaxed LP are actually integer in most cases. In over 300,000 numerical experiments summarized in the following section, more than 95% experiments resulted in integer optimal solutions. In the rest 5% cases, the following rounding method (see Algorithm 2) was applied which always generated integer solutions while maintaining the same optimal values for the objective functions of Eq. (15).

When the optimal control generated by the Linear Program is non-integer, the following empirical fact is always observed:

4.1.1. Empirical fact

If there exists some time t_0 and $i_0 \in Q_s$ such that $u_{t_0}^{i_0,k}$ is non-integer, then it is always true that there exist Δt_1 and Δt_2 such that

$$\sum_{t=t_0-\Delta t_1}^{t_0+\Delta t_2} \sum_{i \in \mathbb{I}} u_t^{i,k} \in \mathbb{Z}_+ \tag{32}$$

where $u_t^{i,k}$ is the delay control in cell (k, i) at time t and

$$\mathbb{I} = \{i : (i,k) \in Q_s \quad \text{and} \quad i_0 - \Delta t_1 \leqslant i \leqslant i_0 + \Delta t_2\}.$$

This can be interpreted as follows: we can always find (a) different control(s) $u_t^{i,k}$, which is (are) applied to the flights on the same path (path k in this case, going from the same origin to the same destination) in the same sector s, such that the overall control is integer. With integer control u and integer inputs f, the state x is always integer using the dynamic Eq. (4). The investigation of the reason for this fact is ongoing work.

Based on the fact above, a rounding method is proposed in Algorithm 2 to round the optimal control output u to an integer value: for each non-integer $u_{t_0}^{i_0,k}$ on a link in sector s, find Δt_1 and Δt_2 such that Eq. (32) holds. Each non-integer control $u_t^{i,k}$ ($t_0 - \Delta t_1 \leqslant t \leqslant t_0 + \Delta t_2$ and $i \in \mathbb{I}$) will be rounded by taking the nearest integer below the sum of its current value and the residual of the control values that have been rounded (see Algorithm 2 for details).

Algorithm 2. Rounding algorithm.

```
Inputs:
  Sector $s$ on path $k$.
  $\overline{\mathcal{U}} := \emptyset$ (empty set).
Start: Find the smallest $t_1$ such that $u_{t_1}^{i_1,k}$ is non-integer
    for some $i_1 \in Q_s$.
  $\mathcal{U} := \emptyset$ (empty set).
  Initial $u_{\text{overall}}$: = 0.
  Initial $u_{\text{offset}}$: = 0.
  $u_{\text{overall}} := u_{t_1}^{i_1,k}$.
  $\mathcal{U} := \mathcal{U} \cup \{u_{t_1}^{i_1,k}\}$.
  $t_2$: = $t_1$.
repeat
  $t_2$: = $t_2$ + 1.
  Find $i_2 > i_1$and $i_2 \in Q_s$ such that
    $u_{t_2}^{i_2,k}$ is non-integer.
  $\mathcal{U} := \mathcal{U} \cup \{u_{t_2}^{i_2,k}\}$.
  $u_{\text{overall}} := u_{\text{overall}} + u_{t_2}^{i_2,k}$.
until $u_{\text{overall}}$ is an integer.
repeat
  $u_t^{i,k} :=$ first element in $\mathcal{U}$.
  $\mathcal{U} := \mathcal{U} \setminus u_t^{i,k}$.
  $\lfloor u_t^{i,k} \rfloor :=$ floor $\left(u_t^{i,k} + u_{\text{offset}}\right)$.
  $\overline{\mathcal{U}} := \overline{\mathcal{U}} \cup \{\lfloor u_t^{i,k} \rfloor\}$.
  $u_{\text{offset}} := u_{\text{offset}} + u_t^{i,k} - \lfloor u_t^{i,k} \rfloor$.
  $u_{\text{overall}} := u_{\text{overall}} - \lfloor u_t^{i,k} \rfloor$.
until $u_{\text{overall}}$ = 0.
Output: $\overline{\mathcal{U}}$.
```

An example of the rounding method is illustrated in Table 2. The number to the left of an arrow is an optimal control u output from the LP, while the value to the right of the arrow is the rounding result using Algorithm 2. The explanation of the example follows: the delay control action with $u_1^2 = 1.3$ and $u_2^3 = 1.7$ is equivalent to $\lfloor u_1^2 \rfloor = 1$ and $\lfloor u_2^3 \rfloor = 2$.

5. Results

The dual decomposition method was implemented in C++, with subproblems (24) solved using ILOG CPLEX Concert (ILOG CPLEX, 2009). A 2-h TFM problem was solved for the whole continental NAS in the United States.

Fig. 4 shows the amount of delay control in each sector using the dual decomposition method at different times. Sectors with a larger controlled count are colored with a darker color.

Fig. 5 shows a comparison between sector counts for three sectors. For each of them, the counts are displayed for an uncontrolled scenario, in which aircraft are flying according to their flight plan (when the controls u in system (3) are set to be zero) and a controlled scenario (when the controls u are optimized by the dual decomposition method). The sector capacities are also represented in the figures as a reference. As can be seen from the solution, by generating an optimal delay allocation in the NAS, the dual decomposition algorithm minimizes the total travel time of the flights in the entire airspace while respecting sector capacity constraints. For example in Fig. 5, the capacity of sector ZOB26 (a high altitude sector in Cleveland ARTCC) is 18; when no delay control is applied, the number of flights in ZOB26 is above 18 after 52 min and can reach 24 (at 99 min), which exceeds the sector capacity (18 flights at a time), while the number of flights stays below the sector capacity all time when delay control is applied. Fig. 5 also shows the situation for sector ZOB29, a neighbor sector of ZOB26 in Cleveland ARTCC, whose capacity is 17. When no delay control is applied, the number of flights in ZOB29 is always below the capacity (under-utilized). However, with an optimal delay control, the dual decomposition algorithm allocates delays in ZOB29, increasing its sector load, which reaches the sector capacity at 35–37 min.

One of the outcomes of the optimization algorithm can be qualitatively described as follows: congestion in sectors will be reduced by allocating delays in under-utilized sectors (usually in under-utilized neighboring sectors).

Other experiments have been conducted with different weights of $c(i, j)$ associated with the cost of flying through cell (i, j), to assess the distribution of delays and the resulting fairness issues in the TFM application. Fig. 6 shows the sector counts

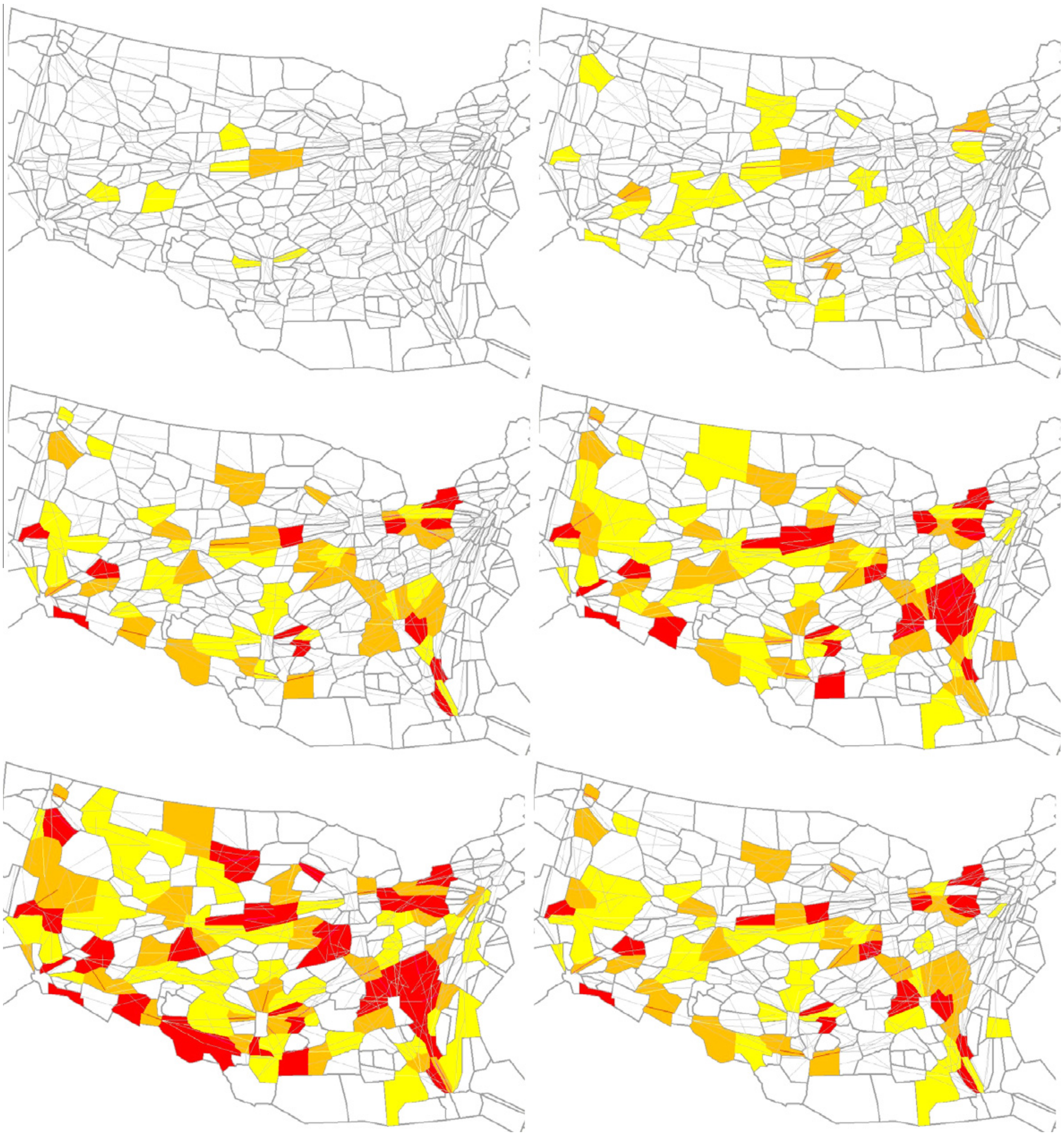

Fig. 4. Solution 19 (top left), 39 (top right), 59 (mid left), 79 (mid right), 99 (bottom left), 119 (bottom right) minutes after the start time, with delay control computed by the dual decomposition algorithm. Shading of sectors indicates aircraft occupancy in each sector with control.

in sector ZLA16 with the costs of flying through cells in ZLA16 equals one, two and four respectively, while keeping the costs of cells in all other sectors remaining one. As an illustration of the distribution of delays in other sectors, Fig. 7 shows the aircraft counts in sector ZLA34, which is a sector adjacent to ZLA16. As can be seen from Fig. 7, more flights are delayed in ZLA34 due to smaller cost for delays in this sector as compared with that in ZLA16. With larger costs for delays in ZLA16 (e.g., increased from 2 to 4), there are more delays caused in ZLA34. It is our future work to assess the global distribution of delays and to investigate the design of appropriate delay costs in different sectors to include fairness issues in the study. Computation wise, different costs did not change the computational time to run the optimization, which remains around 2.5 min for a 2-h TFM problem.

The computation is done on a Dell T1500 workstation, with a 2.8 GHz 8-processor CPU and 16G RAM, running RedHat Enterprise Linux 5 operating system. Note that in light of the fact that the dual decomposition method is highly suitable for parallel computing, a parallel computing platform is designed to improve the computational efficiency. Fig. 8 shows the run time improvement by deploying different numbers of processes. Observed from this figure, the more processes are executed simultaneously, the more run time it saves. The run time reduction decreases as the number of processes

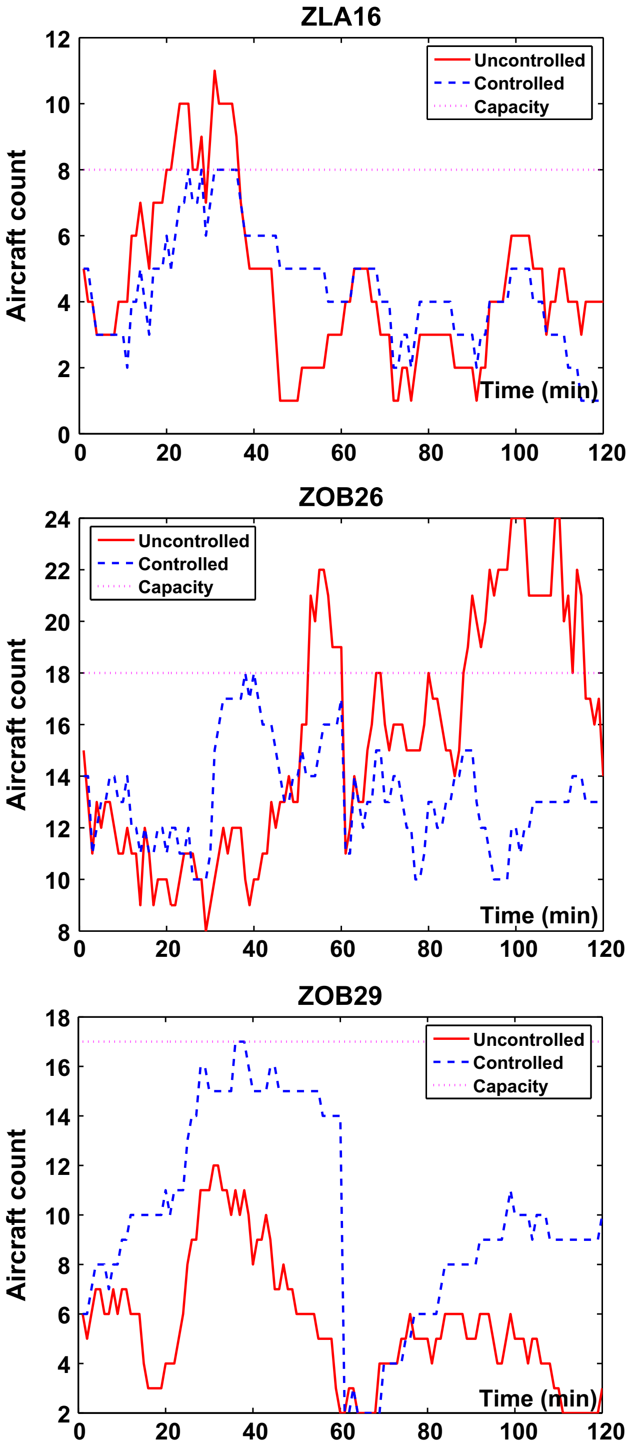

Fig. 5. Comparison of controlled and uncontrolled sector counts in Sectors ZLA16, ZOB26 and ZOB 29.

approaches eight (which is the number of processors in the workstation) and undergoes slight changes with more processors. This is due to the overhead for synchronization and job distribution. A more general illustration of the parallel computing platform is presented in Cao and Sun (in press). The computation presented in this paper is done on a single computer with 8 processes, which will be implemented on a more general parallel computing platform described in Cao and Sun (in press) in the future. The computational time (using eight processes in a single workstation) for a 2-h TFM problem is about 2.5 min. Fig. 9 summarizes the average and maximum computational times for at least 10,000 experiments in each 'day': the average/maximum run time for day *i* is the average/maximum computational time for at least 10,000 experiments from the *i*th day of every month in a year (from October 1st of 2004 to September 30 of 2005). For a particular day in a particular month, approximately 100 scenarios (100 different 2-h intervals) are considered, and each scenario is run 10 times. For example, the average computational time for day 3 in Fig. 9 is 41 min, which is the average computational time for a 2-h

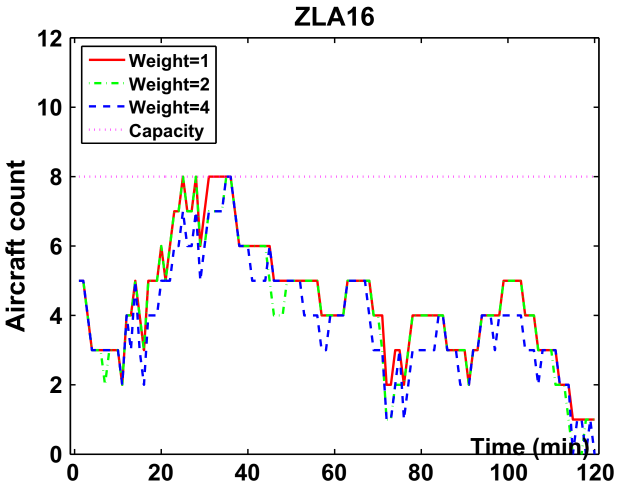

Fig. 6. Sector counts in ZLA16 when different cost weights are used.

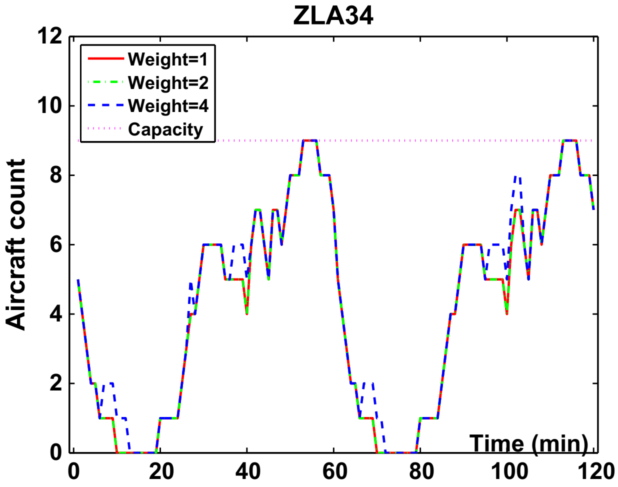

Fig. 7. Sector counts in ZLA34 when different cost weights are used for the neighboring sector ZLA16. The weight in ZLA34 is one for all the cases. Numbers in legend are the weights used in ZLA16.

TFM problem from 1200 scenarios taken from the 3rd day of each month in a year (October 3rd of 2004, November 3rd of 2004, etc., till September 3rd of 2005). For each day, e.g., September 3rd of 2005, 100 time intervals are chosen for the 2-h optimization time horizon, which include 8:00 AM–10:00 AM EDT (12:00–14:00 UTC), 8:10 AM–10:10 AM EDT, 8:45 AM–10:45 AM EDT, for example.

With a fixed network model of the NAS and a fixed optimization time horizon (2 h in our examples), we have a fixed number of variables and constraints: approximately five billion variables and eight billion constraints. The size of a subproblem mainly depends on the length of the associated path. Typically a subproblem involves 6000 variables and 10,000 constraints. The computational time does not significantly change with different scenarios (amount of flights, level of sector capacities, different days, etc.).

As a comparison with decomposition-based TFM algorithms in literature, particularly the most recent study in Rios and Ross (2010), similar traffic scenarios are tested and summarized in Table 1. The application of the dual decomposition method differs from that of the Dantzig–Wolfe decomposition method in Rios and Ross (2010) in the sense that they are based on different models: the basis for the dual decomposition method is the CTM(L) (Sun and Bayen, 2008), an aggregate model whose dimension (number of variables and constraints) and complexity do not change with the number of flights, while the base model for the Dantzig–Wolfe decomposition method is the seminal Bertsimas and Patterson model (Bertsimas and Stock Patterson, 1998), which is Lagrangian. From Table 1, it can be seen that the dual decomposition method is comparable to the Dantzig–Wolfe decomposition method in terms of run time; most importantly, we can conclude that decomposition methods provide an opportunity to accelerate the computation for TFM, which makes a real-time TFM platform possible.

The CTM(L) and the dual decomposition method for solving NAS-wide TFM problems have been integrated with FACET (Bilimoria et al., 2001) by Metron Aviation (2010). The output from the dual decomposition method based on CTM(L) is

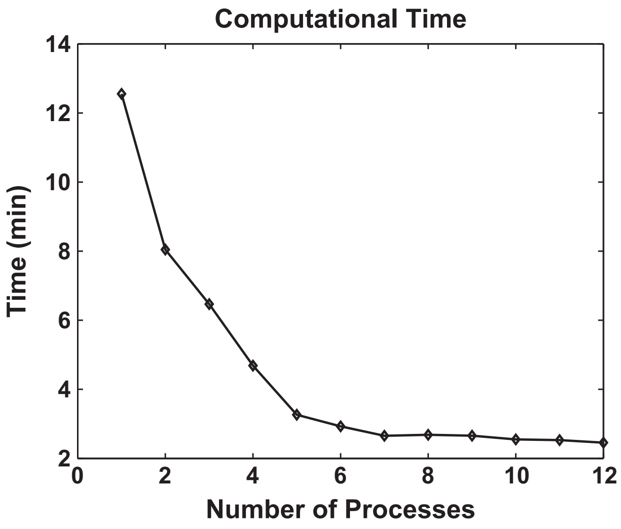

Fig. 8. Computational time with different number of processes.

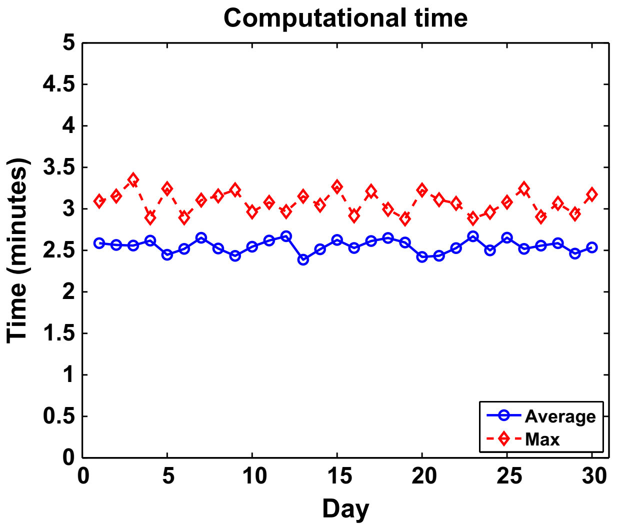

Fig. 9. Computational time of the dual decomposition method for 2-h TFM problems.

Table 1
A comparison of runtime between the dual decomposition method and a Dantzig–Wolfe decomposition method (Rios and Ross, 2010).

	Dual decomposition	Dantzig–Wolfe decomposition
Planning time horizon (min)	120	120
Number of flights	7956	8505
Reduced sector capacity (%)	90	90
Run time (s)	149	101

Table 2
An example of the rounding algorithm.

Time	Cell			
	1	2	3	4
1	0 → 0	1.3 → 1	0 → 0	0 → 0
2	0 → 0	0 → 0	1.7 → 2	0 → 0

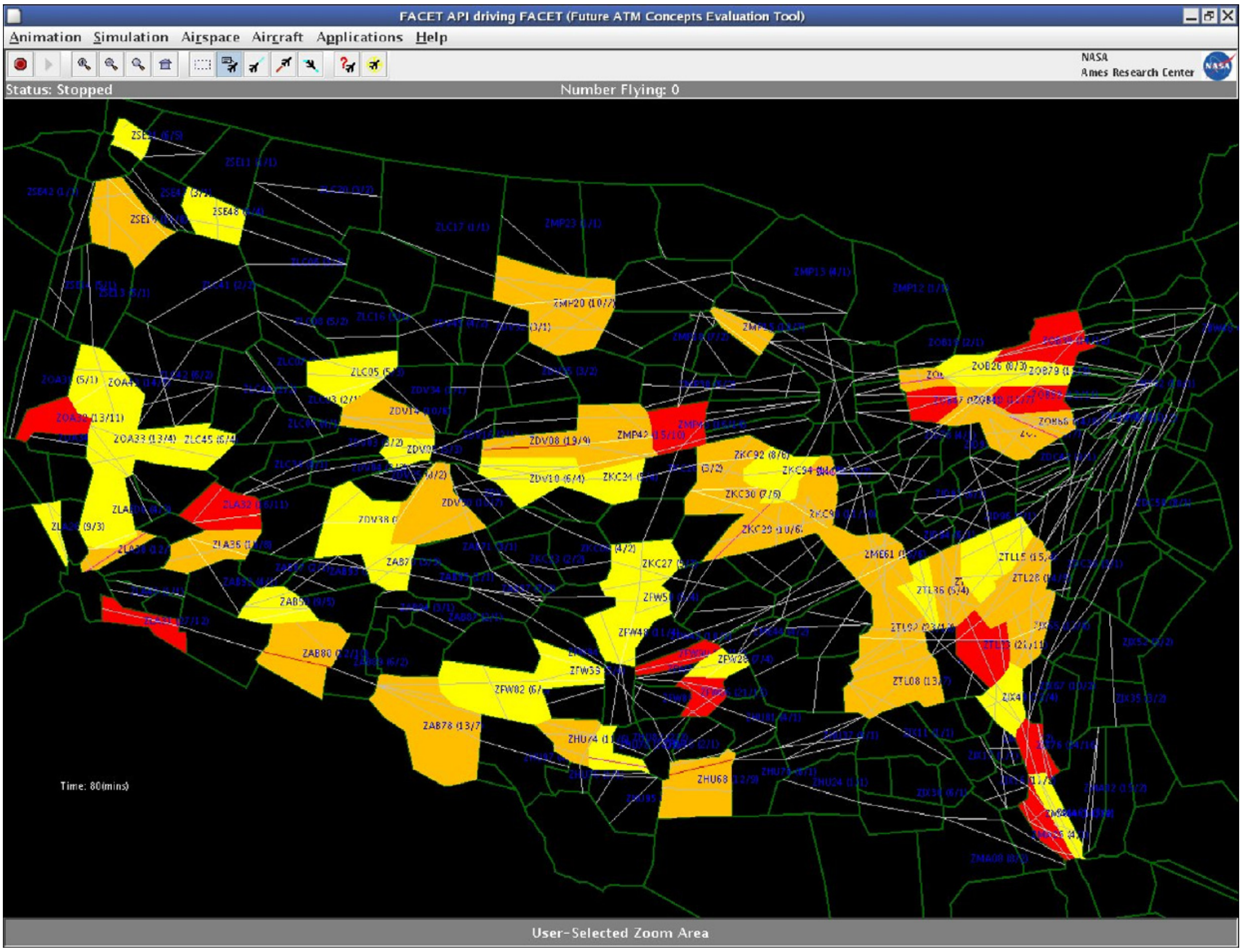

Fig. 10. A snapshot of FACET integrated with CTM(L) and dual decomposition method for solving NAS-wide TFM problems.

optimal delay control for groups of flights at specific times in specific sectors. FACET then assigns delay controls (such as change of speed, holding patterns) to individual flights associated with the groups of flights in the CTM(L) (Hoffman et al., 2008), such that the TFM requirements defined in the optimization problem can be achieved. Fig. 10 shows a snapshot of the integration of this work in FACET. The details of the implementation of this algorithm in FACET will be the focus of a separate publication, joint with Metron, in a systems journal.

6. Conclusion

In this article, a linear time-varying aggregate traffic flow model based on multicommodity flow network is developed and implemented. Using the model, an optimization framework is proposed for strategic Traffic Flow Management problems, which are formulated as Integer Programs. The generic Integer Program is relaxed to a Linear Program for computational efficiency, which is solved by a dual decomposition method developed specifically for the multicommodity flow problem. The program is implemented for the entire National Airspace System in the continental United States, which results in an implementation with billions of variables and constraints. Optimal delay controls are obtained based on the resulting optimization solutions, which are tractable at the scale of the United States. The work presented in this article has been integrated with NASA's Future ATM Concepts Evaluation Tool in a first step towards developing an operational tool available for practitioners.

Acknowledgments

This work was supported by NASA under Task Order TO.048.0.BS.AF, NASA NRA Subtopic 4, and startup fund from College of Engineering, Purdue University. The authors want to acknowledge Carlos Daganzo for the insightful discussions about his Cell Transmission Model and his suggestion to name the model presented in this study the Large-capacity Cell Transmission Model. The authors are grateful to Banavar Sridhar for his vision on aggregate traffic flow modeling, and fruitful conversations on air traffic control. The authors are thankful to Shon Grabbe and Kapil Sheth for their help with ASDI/ETMS data and

FACET. Gano Chatterji is gratefully acknowledged for his suggestions regarding the models of the National Airspace System. Laurent El Ghaoui helped the authors with large scale optimization and the dual decomposition method, and is warmly acknowledged. Metron Aviation contributed to the integration of our algorithm to the software FACET developed at NASA Ames. The authors are grateful to Larry Hogle for helping with the collaboration between UC Berkeley and NASA Ames through the UARC.

The authors are very thankful to the reviewers of this article for their thorough reading and constructive comments. In particular, the reviewers' comments for corrections of the dual decomposition method are highly appreciated.

Appendix A. CTM(L) and TFM with airport

A.1. Model

The CTM(L) can be extended to include airports as part of the model. This allows the use of ground holding, which is less expensive than airborne delays. The network modeling of the air traffic can be extended as follows.

A.1.1. Vertices (nodes)

For any airport, for departure flights a_o (which will be referred as an "origin airport"), a vertex denoted as $v_{\{\text{ground},a_o\}}$ represents the entry point used by flights entering the network and waiting for departure, while a vertex $v_{\{a_o\}}$ represents the entry point used by departure flight going from airport a_o to a *low* sector. Similarly, vertices $v_{\{a_d,\text{ground}\}}$ and $v_{\{a_d\}}$ are defined for an airport for arrival flights a_d (which will be referred as a "destination airport").

A.1.2. Links (edges)

Let $V_i^{\text{out}} = \{v_1, v_2, \ldots, v_{|V_i^{\text{out}}|}\}$ be the set of vertices representing exit points from sector $s_i \in S$ to the neighboring sectors. The set of vertices which represents entry points from neighboring sectors to sector s_i is denoted by $V_i^{\text{in}} = \{v_1, v_2, \ldots, v_{|V_i^{\text{in}}|}\}$. For each origin airport a_o in sector s_i, *directed links* are created: one from vertex $v_{\{\text{none},a_o\}}$ to vertex $v_{\{a_o\}}$ and one from vertex $v_{\{a_o\}}$ to each vertex $v_k \in V_i^{\text{out}}$. Similarly, *directed links* from each vertex $v_k \in V_i^{\text{in}}$ to vertex $v_{\{a_d\}}$ and one *directed link* from vertex $v_{\{a_d\}}$ to vertex $v_{\{a_d,\text{none}\}}$ are created for any destination airport a_d in sector s_i. For the rest of this work, the term low-altitude link refers to a directed link from an airport to a high-altitude sector, and vice versa. The directed link included between an airport and the ground is called a *queuing link*.

The aggregated flight time along a low-altitude link is estimated using the great circle distance and an estimation of flight speed. For each single flight, this distance is computed between two points: one from the flight coordinates in the most recent recorded ASDI/ETMS data and one from the physical location of the airport.

Fig. 11 shows an interface between an origin airport a_o and high-altitude sectors.

A.1.3. Extended network

Flights are now clustered based on their origin-destination airport pairs. The network is still defined as an aggregation of the trees for all destination airports, instead of destination sectors. Based on the extended network, the CTM(L) model now includes low-altitude links and airports queuing links.

A.2. Ground delay

Using the extended CTM(L), the problem of minimizing the total flight travel time in the NAS can now be extended to include departure airports. The optimization formulation remains unchanged as in Section 3.2, but the equation which

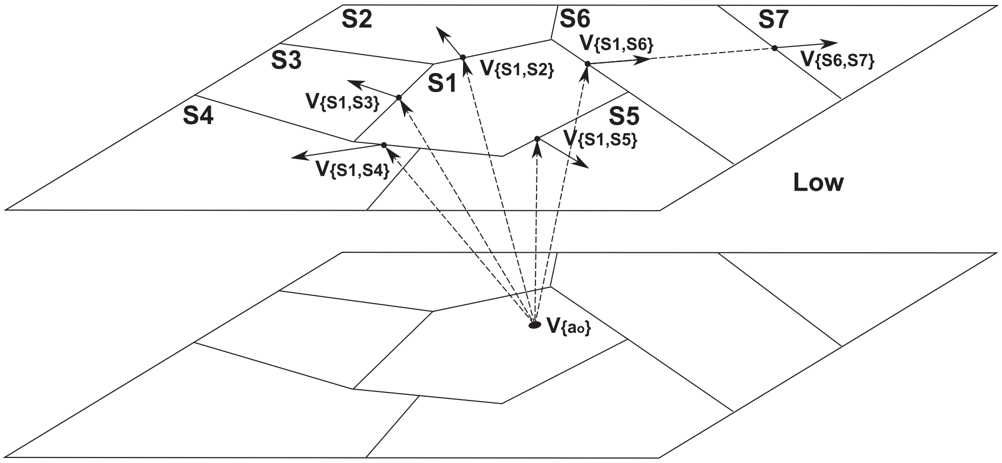

Fig. 11. An illustration of vertices and links interface for an origin airport.

encodes the dynamics of the system now includes departure airport queuing system. The queuing departure link is defined by the state vector $x_a^{\text{dep}}(t) = [x_a^1(t); \ldots; x_a^{m_a}(t)]$, whose elements represent the corresponding aircraft counts in each cell of the departure airport queuing link a at time step t, and m_a is the number of cells in the link. The number of departure flights from airport a at time $t + 1$ is equal to $x_a^{m_a}(t)$ unless delay control action is applied.

Each origin-destination pair in the network corresponds to a *path*, which consists of queuing links, low-altitude links and en route links between these nodes. The same rules of a multicommodity network flow model are applied, and if two or more paths have a common queuing link or low-altitude link, this link will be duplicated. The forcing input, $f(t)$, is now the entry count of departure aircraft, m_a minutes before their departure time from airport a.

The cost c includes the cost of ground holding per plane per time period (c_g) and the cost of airborne holding per plane per time period (c_a). In the rest of this work, the cost is defined with an air to ground delay cost ratio equals 3.

A.3. Destination airport capacities

Congestion at a destination airport is defined by the difference between the number of arrival flights within a 15-min interval and the capacity of the current runway configuration. The dual decomposition method, developed in Section 4, is now adapted for the optimization for traffic flow for the entire network including airports.

The additional notations are defined below.

- $\mathcal{A} = \{a_1, a_2, \ldots, a_{\mathcal{A}}\}$: set of destination airports in the NAS. $|\mathcal{A}|$ denotes total number of airports in the NAS.
- $\mathbb{A} = \{1, 2, \ldots, |\mathcal{A}|\}$: set of destination airport indices.
- $C_a^{\delta t}(\theta), a = 1, 2, \ldots, |\mathcal{A}|$: airport capacity of the current runway configuration of airport a for a δt-minute interval from t to $t + \delta t$. Denote $C_{\mathbb{A}}^{\delta t}(\theta) = [C_1^{\delta t}(\theta); C_2^{\delta t}(\theta); \ldots; C_{|\mathcal{A}|}^{\delta t}(\theta)]$, and $C_{\mathbb{A}}^{\delta t} = [C_{\mathbb{A}}^{\delta t}(1); C_{\mathbb{A}}^{\delta t}(2); \ldots; C_{\mathbb{A}}^{\delta t}(\Theta)]$.
- $Q_a \subset E$: set of cells in airport $a \in \mathcal{A}$. $(j, k) \in Q_a$ means the jth cell on path k is in Q_a.
- $\Theta = T_1, T_2, \ldots, T_{|\Theta|}$: δt-minute time interval, given by: $T_\theta = [\theta \cdot \delta t + 1, \ldots, (\theta + 1) \cdot \delta t]$. $|\Theta|$ denotes total number of δt-minute for ground delays.
- $\vartheta = 1, 2, \ldots, |\Theta|$: set of δt-minute time interval indices.
- Slack variables $Z_a^k(t)$, $a = 1, \ldots, |\mathcal{A}|$, $k = 1, \ldots, |K|$ representing the number of aircraft in airport a on path k at time t, defined by: $Z_a^k(t) = \sum_{(j,k)\in Q_a} x(k, j, t)$, where (j, k) is the jth cell on path k and Q_a is the set of links for airport a. Let $Z_{\mathbb{A}}^k(t)$ denote the aggregation of slack variables $Z_a^k(t)$ on path k at time $t : Z_{\mathbb{A}}^k(t) = [Z_1^k(t); Z_2^k(t); \ldots; Z_{|\mathcal{A}|}^k(t)]$, and $Z_{\mathbb{A}}^k = [Z_{\mathbb{A}}^k(1); Z_{\mathbb{A}}^k(2); \ldots; Z_{\mathbb{A}}^k(T)]$, slack variables associated with path k. Let $Z_{\mathbb{A}}(t)$ denote the aggregation of all slack variables at time t: $Z_{\mathbb{A}}(t) = [Z_{\mathbb{A}}^1(t); Z_{\mathbb{A}}^2(t); \ldots; Z_{\mathbb{A}}^{|K|}(t)]$. Let $Z_{\mathbb{A}} = [Z_{\mathbb{A}}(1); Z_{\mathbb{A}}(2); \ldots; Z_{\mathbb{A}}(T)]$ be the aggregation of all slack variables.

A new constraint is included in the problem of minimizing the total travel time of the flights in the NAS. This constraint is formulated as follows:

$$\sum_{t\in T_\theta} \sum_{(i,j)\in Q_a} x_t^{i,j} \leqslant C_a^{\delta t}(\theta), \qquad a \in \mathbb{A}, \quad \theta \in \vartheta$$

The optimization problem is formulated as follows:

$$\min_{x,u} \quad \sum_{t=0}^{T} c^T x_t, \tag{33}$$

$$\text{s.t.} \quad x_0 = B_2 f', \tag{34}$$

$$x_{t+1} = Ax_t + B_1 u_t + B_2 f_t, \qquad t \in \mathbb{T}_0, \tag{35}$$

$$\sum_{(i,j)\in Q_s} x_t^{i,j} \leqslant C_s(t), \qquad s \in \mathbb{S}, \quad t \in \mathbb{T}, \tag{36}$$

$$\sum_{t\in T_\theta} \sum_{(i,j)\in Q_a} x_t^{i,j} \leqslant C_a^{\delta t}(\theta), \qquad a \in \mathbb{A}, \quad \theta \in \vartheta, \tag{37}$$

$$u \leqslant x \tag{38}$$

$$x \subset \mathbb{Z}_+ \tag{39}$$

$$u \subset \mathbb{Z}_+ \tag{40}$$

Similar constraints as in (18)–(20) are enforced to make sure every scheduled flight should will land at its destination airport. Solving this Integer Program follows the same scheme as discussed in Section 4.

A.4. Results

The extended framework was implemented on the same platform as in Section 5. A 2-h TFM problem was solved for the whole continental NAS in the United States, with arrival capacity constraints in three major destination airports. To reduce

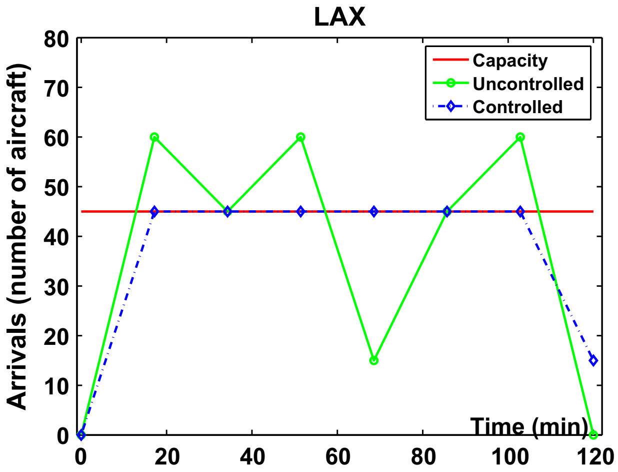

Fig. 12. Comparison of controlled and uncontrolled arrivals at the Los Angeles Airport (LAX).

the number of variables and constraints, $\delta t = 15$ min is chosen as the update time interval. The computational time of solving this optimization problem is less than 60 min.

The airport capacity constraint is fixed as $C_a^{\delta t}(\theta) = 45, \theta \subset \vartheta$, which means that the *Airport Arrival Rate* (AAR) accepted by the airport is up to 45 aircraft per δt minutes. In this study, three airports have a restricted arrival capacity constraint, which are Los Angeles (LAX), San Francisco (SFO) and Boston Logan International Airport (BOS).

Fig. 12 shows a comparison between the real and the optimal aircraft arrival counts at LAX. The airport arrival capacity constraint is also represented as a reference. For example the schedule arrivals for LAX airport for the 15 min interval from $t = 1$ to $t = 15$ min are 60 flights, which exceeds the arrival capacity. The optimal solution gives an optimal number of arrivals equal to the airport arrival capacity from $t = 20$ to $t = 100$. This solution makes perfect physical sense, i.e., it shows that an optimal integrated high altitude – airport problem spreads capacity usage between the ground and the air, in particular by maximization of airport runway usage.

References

Ahuja, R.K., Magnati, T.L., Orlin, J.B., 1993. Network Flows: Theory Algorithms and Application. Prentice Hall, Upper Saddle River, NJ.

Andreatta, G., Brunetta, L., 1998. Multi-airport ground holding problem: a computational evaluation of exact algorithms. Operations Research 46 (1), 57–64.

Andreatta, G., Brunetta, L., Guastalla, G., 1995. Multi-airport ground holding problem: a heuristic approach based on priority rules. In: Bianco, L., Dell'Olmo, P., Odoni, A. (Eds.), Modeling and Simulation in Air Traffic Management. Springer Verlag, Berlin, Germany.

Anonymous, 2005. Terminal Area Forecast Summary, Fiscal Year 2004–2020. Tech. Rep. FAA-APO-05-1, US Department of Transportation, Federal Aviation Administration.

Ball, M., Hoffman, R., Lovell, D., Mukherjee, A., 2005. Response Mechanisms for dynamic air traffic flow management. In: Proceedings of the 6th USA/Europe ATM 2005 R&D Seminar. Baltimore, MD, June.

Ball, M., Hoffman, R., Mukherjee, A., 2010. Ground delay program planning under uncertainty based on the ration-by-distance principle. Transportation Science 44 (1), 1–14.

Ball, M., Lulli, G., 2004. Ground delay programs: optimizing over included flight set based on distance. Air Traffic Control Quarterly 12 (1), 1–25.

Bayen, A.M., Raffard, R.L., Tomlin, C.J., 2006. Adjoint-based control of a new Eulerian network model of air traffic flow. IEEE Transactions on Control Systems Technology 14 (5), 804–818.

Bazaraa, M., Sherali, H., Shetty, C., 2006. Nonlinear Programming: Theory and Algorithms, third ed. Wiley-Interscience.

Bertsekas, D., 2002. Nonlinear Programming, second ed. Athena Scientific, Belmont, MA.

Bertsimas, D., Lulli, G., Odoni, A., 2008. The air traffic flow management problem: an integer optimization approach. In: Lodi, A., Panconesi, A., Rinaldi, G. (Eds.), Integer Programming and Combinatorial Optimization. Lecture Notes in Computer Sciences, vol. 5035. Springer-Verlag, Berlin, Heidelberg.

Bertsimas, D., Stock Patterson, S., 1998. The air traffic flow management problem with enroute capacities. Operations Research 46 (3), 406–422.

Bilimoria, K., Sridhar, B., Chatterji, G., Sheth, K., Grabbe, S., 2001. FACET: future ATM concepts evaluation tool. Air Traffic Control Quarterly 9 (1), 1–20.

Bloem, M., Sridhar, B., 2008. Optimally and equitably distributing delays with the aggregate flow model. In: IEEE/AIAA 27th Digital Avionics Systems Conference (DASC). St. Paul, Minnesota, October, pp. 3.D.4-1–3.D.4-14.

Bolczak, C., Hoffman, J., Jensen, A., Trigeiro, W., 1997. National Airspace System Performance Measurements: Overview. Tech. Rep. 97W0000035, MITRE.

Boyd, S., Vandenberghe, L., 2004. Convex Optimization. Cambridge University Press, New York, NY.

Bradford, S., Knorr, D., Liang, D., 2000. Performance measures for future architecture. In: Proceedings of the 3rd USA/Europe ATM 2000 R&D Seminar, Napoli, Italy, June.

Bratu, S., Barnhart, C., 2005. An analysis of passenger delays using flight operations and passenger booking data. Air Traffic Control Quarterly 13 (1), 1–29.

Callaham, M., DeArmon, J., Cooper, A., Goodfriend, J., Moch-Mooney, D., Solomos, G., 2001. Assessing NAS performance: normalizing for the effects of weather. In: Proceedings of the 4th USA/Europe ATM 2001 R&D Seminar, Santa Fe, NM, December.

Cao, Y., Sun, D., in press. A Parallel Computing Platform for Air Traffic Flow Management. In: American Control Conference, San Francisco, CA.

Chatterji, G., Sridhar, B., 2005a. National airspace system delay estimation using weather weighted traffic counts. In: Proceedings of the AIAA Guidance, Navigation and Control Conference and Exhibit, San Francisco, CA, August.

Chatterji, G., Sridhar, B., 2005b. Some properties of the aggregate flow model of air traffic. In: AIAA 5th ATIO and 16th Lighter-Than-Air Sys Tech. and Balloon Systems Conferences, Arlington, Virginia, September.

Chiang, M., Low, S., Calderbank, A., Doyle, J., 2007. Layering as optimization decomposition: a mathematical theory of network architectures. Proceedings of the IEEE 95 (1), 255–312.

Churchill, A., Lovell, D., Ball, M., 2007. Examining the temporal evolution of propagated delays at individual airports: case studies. In: Proceedings of the 7th USA/Europe ATM 2007 R&D Seminar, Barcelona, Spain, July.

Churchill, A., Lovell, D., Ball, M., 2010. Flight delay propagation impact on strategic air traffic flow management. Transportation Research Record 2177, 105–113.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C., 2002. Introduction to Algorithms, second ed. Prentice Hall, Upper Saddle River, NJ.

Daganzo, C., 1994. The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. Transportation Research Part B 28 (4), 269–287.

Daganzo, C., 1995. The cell transmission model, part II: network traffic. Transportation Research Part B 29 (2), 79–93.

Dantzig, G., Wolfe, P., 1960. Decomposition principle for linear programs. Operations Research 8 (1), 101–111.

Dell'Olmo, P., Lulli, G., 2003. A dynamic programming approach for the airport capacity allocation problem. IMA Journal of Management Mathematics 14 (3), 235–249.

Erzberger, H., 1992. CTAS: Computer Intelligence for Air Traffic Control in the Terminal Area. Tech. Rep. NASA TM 103959, Ames Research Center.

Gilbo, E., 1993. Airport capacity: representation, estimation, optimization. IEEE Transactions on Control Systems Technology 1 (3), 144–154.

Gosling, G., 1999. Aviation System Performance Measures. NEXTOR Working Paper UCB-ITS-WP-99-1.

Grabbe, S., Sridhar, B., Mukherjee, A., 2007. Central east pacific flight scheduling. In: AIAA Conference on Guidance, Navigation, and Control Conference and Exhibit, Hilton Head, SC, August.

Grabbe, S., Sridhar, B., Mukherjee, A., 2009. Sequential traffic flow optimization with tactical flight control heuristics. AIAA Journal of Guidance, Control, and Dynamics 32 (3), 810–820.

Helme, M., 1992. Reducing air traffic delay in a space-time network. In: IEEE International Conference on Systems, Man and Cybernetics, vol. 1. pp. 236–242.

Hoffman, R., Ball, M., 2000. A comparison of formulations for the single airport ground holding problem with banking constraints. Operations Research 48 (4), 578–590.

Hoffman, R., Sun, D., Clinet, A., Augustine, S., Burke, J., Viswanathan, R., Bayen, A., 2008. Integration of an aggregate flow model with a traffic flow simulator. In: AIAA Conference on Guidance, Navigation, and Control Conference and Exhibit, Honolulu, HI, August.

Idris, H., Evans, A., Vivona, R., Krozel, J., Bilimoria, K., 2006. Field observations of interactions between traffic flow management and airline operations. In: 6th AIAA Aviation Technology, Integration, and Operations Conference, Wichita, KS, September.

ILOG CPLEX, 2009. <http://www.ilog.com/corporate/training/acrobat/CPLEX.pdf> (retrieved 07.04.10).

Joint Planning and Development Office, 2009. NextGen Concept of Operations, version 2.0. <http://www.jpdo.gov/library/NextGen_v2.0.pdf> (retrieved 07.04.10).

Kelly, F., Maulloo, A., Tan, D., 1997. Rate control for communication networks: shadow prices, proportional fairness and stability. Journal of the Operational Research Society 49 (3), 237–252.

Klein, A., 2007. NAS/ATM performance indexes. In: Proceedings of the 7th USA/Europe ATM 2007 R& D Seminar, Barcelona, Spain, July.

Kopardekar, P., Green, S., 2005. Airspace restriction planner for sector congestion management. In: AIAA 5th ATIO and 16th Lighter-Than-Air Sys Tech. and Balloon Systems Conferences, Arlington, Virginia, September.

Kopardekar, P., Prevot, T., Jastrzebski, M., 2009. Traffic complexity measurement under higher levels of automation and higher traffic densities. Air Traffic Control Quarterly 17 (2), 125–148.

Laskey, K., Xu, N., Chen, C., 2006. Propagation of delays in the national airspace system. In: Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence. Cambridge, MA.

Lighthill, M.J., Whitham, G.B., 1956. On kinematic waves. II. A theory of traffic flow on long crowded roads. Proceedings of the Royal Society of London 229 (1178), 317–345.

Lindsay, K., Boyd, E., Burlingame, R., 1993. Traffic flow management modeling with the time assignment model. Air Traffic Control Quarterly 1 (3), 255–276.

Lulli, G., Odoni, A., 2007. The european air traffic flow management problem. Transportation Science 41 (4), 431–443.

Menon, P.K., Sweriduk, G.D., Bilimoria, K., 2002. A new approach for modeling, analysis and control of air traffic flow. In: AIAA Conference on Guidance, Navigation, and Control, Monterey, CA, August.

Metron Aviation, 2010. <http://www.metronaviation.com> (retrieved 07.04.10).

Navazio, L., Romanin-Jacur, G., 1998. The multiple connections multi-airport ground holding problem: models and algorithms. Transportation Science 32 (3), 268–276.

Nolan, M.S., 2003. Fundamentals of Air Traffic Control, fourth ed. Brooks Cole, Reading, MA.

Odoni, A., 1987. The flow management problem in air traffic control. In: Odoni, A., Bianco, L., Szego, G. (Eds.), Flow Control of Congested Networks. Springer Verlag, Berlin, Germany.

Palomar, D., Chiang, M., 2007. Alternative distributed algorithms for network utility maximization: framework and applications. IEEE Transactions on Automatic Control 52 (12), 2254–2269.

Raffard, R., Tomlin, C., Boyd, S., 2004. Distributed optimization for cooperative agents: application to formation flight. In: Proceedings of the IEEE Conference on Decision and Control, Nassau, Bahamas, December, pp. 2453–2459.

Rhodes, L.S., Rhodes, L., Beaton, E., 2001. CRCT Capabilities Detailed Functional Description. Tech. Rep. 00W0000302, MITRE.

Richards, P.I., 1956. Shock waves on the highway. Operations Research 4 (1), 42–51.

Rios, J., Ross, K., 2008. Solving high-fidelity, large-scale traffic flow management problems in reduced time. In: AIAA Aviation Technology, Integration and Operations Conference. Anchorage, Alaska, September.

Rios, J., Ross, K., 2010. Massively parallel Dantzig–Wolfe decomposition applied to traffic flow scheduling. Journal of Aerospace Computing, Information, and Communication 7 (1), 32–45.

Robelin, C.A., Sun, D., Wu, G., Bayen, A.M., 2006. MILP control of aggregate Eulerian network airspace models. in: Proceedings of the American Control Conference, Minneapolis, MN, June, pp. 5257–5262.

Sherali, H., Smith, J., Trani, A., 2002. An airspace planning model for selecting flights plans under workload, safety, and equity considerations. Transportation Science 36 (4), 378–397.

Sherali, H., Staats, R., Trani, A., 2003. An airspace planning and collaborative decision making model: part I – probabilistic conflicts, workload, and equity consideration. Transportation Science 37 (4), 434–456.

Sherali, H., Staats, R., Trani, A., 2006. An airspace planning and collaborative decision making model: part II – cost model, data considerations, and computation. Transportation Science 40 (2), 147–164.

Shor, N., Kiwiel, K., Ruszcayn̂ski, A., 1985. Minimization Methods for Non-differentiable Functions. Springer-Verlag, New York, NY.

Sridhar, B., Chatterji, G., Grabbe, S., Sheth, K., 2002. Integration of traffic flow management decisions. In: AIAA Conference on Guidance, Navigation, and Control Conference and Exhibit, Monterey, CA, August.

Sridhar, B., Grabbe, S., Mukherjee, A., 2008. Modeling and optimization in traffic flowmanagement. Proceedings of the IEEE 96 (12), 2060–2080.

Sridhar, B., Menon, P.K., 2005. Comparison of linear dynamic models for air traffic flow management. In: 16th IFAC World Congress, Prague, Czech, July.

Sridhar, B., Soni, T., Sheth, K., Chatterji, G., 2006. Aggregate flow model for air-traffic management. AIAA Journal of Guidance, Control, and Dynamics 29 (4), 992–997.

Sridhar, B., Swei, S., 2006. Relationship between weather, traffic and delay based on empirical methods. In: 6th AIAA Aviation Technology, Integration, and Operations Conference. Wichita, KS, September.

Sridhar, B., Swei, S., 2007. Classification and computation of aggregate delay using center-based weather impacted traffic index. In: AIAA 7th Aviation Technology, Integration and Operations (ATIO) Forum. Belfast, Northern Ireland, September.

Sun, D., Bayen, A., 2008. Multicommodity Eulerian–Lagrangian Large-capacity Cell Transmission Model for en route traffic. AIAA Journal of Guidance, Control, and Dynamics 31 (3), 616–628.

Sun, D., Sridhar, B., Grabbe, S., 2009. Traffic flow management using aggregate flow models and the development of disaggregation methods. In: AIAA Conference on Guidance, Navigation, and Control Conference and Exhibit, Chicago, IL, August.

Sun, D., Sridhar, B., Grabbe, S., 2010. Disaggregation method for an aggregate traffic flow management model. AIAA Journal of Guidance, Control, and Dynamics 33 (3), 666–676.

Sun, D., Strub, I., Bayen, A., 2007. Comparison of the performance of four Eulerian network flow models for strategic air traffic management. AIMS Journal on Networks and Heterogeneous Media 2 (4), 569–595.

Sun, D., Yang, S., Strub, I., Bayen, A., Sheth, K., Sridhar, B., 2006. Eulerian trilogy. In: AIAA Conference on Guidance, Navigation, and Control Conference and Exhibit. Keystone, CO, August.

Tan, C., Palomar, D., Chiang, M., 2006. Distributed optimization of coupled systems with applications to network utility maximization. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 5. Chicago, IL, May, pp. 981–984.

Terrab, M., Odoni, A.R., 1993. Strategic flow management for air traffic control. Operations Research 41 (1), 138–152.

US Department of Transportation, Federal Aviation Administration, 2008. Facility Operation and Administration. Order 7210.3V.

Volpe National Transportation Center, 2005. Enhanced Traffic Management System (ETMS). Tech. Rep. VNTSC-DTS56-TMS-002, US Department of Transportation, Cambridge, MA.

Vossen, T., Ball, M., 2005. Optimization and mediated bartering models for ground delay programs. Naval Research Logistics 53 (1), 75–90.

Vossen, T., Ball, M., 2006. Slot trading opportunities in collaborative ground delay programs. Transportation Science 40 (1), 29–43.

Vossen, T., Ball, M., Chen, C., Hoffman, R., 2000. Collaborative Decision Making in Air Traffic Management: Current and Future Research Directions. Tech. Rep., University of Maryland, College Park, MD.

Vossen, T., Ball, M., Hoffman, R., Wambsganss, M., 2003. A general approach to equity in traffic flow management and its application to mitigating exemption bias in ground delay programs. In: Proceedings of the 5th USA/Europe ATM 2003 R&D Seminar, Budapest, Hungary, June.

Vranas, P., Bertsimas, D., Odoni, A.R., 1994. The multi-airport ground holding problem in air traffic control. Operations Research 42 (2), 249–261.

Wang, P., Schaefer, L., Wojcik, L., 2003. Flight connections and their impacts on delay propagation. In: Proceedings of the 22nd IEEE Conference on Digital Avionics Systems. McLean, VA, pp. 5.B.4–5.1–9, October.

Wanke, C., Greenbaum, D., 2007. Incremental probabilistic decision making for en route traffic management. Air Traffic Control Quarterly 15 (4), 299–319.

Waslander, S., Raffard, R., Tomlin, C., 2008a. Market-based air traffic flow control with competing airlines. AIAA Journal of Guidance, Control and Dynamics 31 (1), 148–161.

Waslander, S., Roy, K., Johari, R., Tomlin, C., 2008b. Lump sum markets for air traffic flow control with competitive airlines. IEEE Special Issue on Aviation Information Systems 96 (12), 2113–2130.

Xiao, L., Johansson, M., Boyd, S., 2004. Simultaneous routing and resource allocation via dual decomposition. IEEE Transactions on Communications 52 (7), 1136–1144.

Xu, N., Donohue, G., Laskey, K., Chen, C., 2005. Estimation of delay propagation in the national aviation system using bayesian networks. In: Proceedings of the 6th USA/Europe ATM 2005 R&D Seminar. Baltimore, MD, June.