MODELLING DELAY PROPAGATION WITHIN AN AIRPORT NETWORK

Nikolas Pyrgiotis, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology (Room 35-217, MIT, Cambridge, MA 02139; pyrgios@mit.edu)

Kerry M. Malone, TNO, The Netherlands (P.O. Box 49, 2600 AA Delft, The Netherlands; kerry.malone@tno.nl)

Amedeo Odoni, Department of Aeronautics and Astronautics and Department of Civil and Environmental Engineering, Massachusetts Institute of Technology (Room 33-219, MIT, Cambridge, MA 02139; arodoni@mit.edu)

ABSTRACT

This paper is concerned with the propagation of delays within a large network of major airports. As more airports in the United States and in Europe become more congested, it also becomes increasingly likely that delays at one or more airports will spread to other parts of the network. We describe an analytical queuing and network decomposition model developed to study this complex phenomenon. The Approximate Network Delays (AND) model computes the delays due to local congestion at individual airports and, more important, captures the "ripple effect" that leads to the propagation of these delays. The model operates by iterating between its two main components: a queuing engine (QE) that computes delays at individual airports and a delay propagation algorithm (DPA) that updates flight schedules and demand rates at all the airports in the model in response to the local delays computed by the QE. The QE is a stochastic and dynamic queuing model that treats each airport in the network as a $M(t)/E_k(t)/1$ queuing system.

The AND model is very fast computationally, thus making possible the exploration at a macroscopic level of the impacts of a large number of scenarios and policy alternatives on system-wide delays. It has been implemented for a network consisting of the 34 busiest airports in the continental United States. The model provides insights into the complex interactions through which delays propagate through the network and the often-counterintuitive consequences of these interactions. For example, it shows how the propagation of delays tends to "smoothen" daily airport demand profiles, push more demands into late evening hours, and result in local delays which are smaller than what they would have been with the original demand profiles. Such phenomena are especially evident at hub airports. It is also shown that, at hub airports, some flights may benefit considerably (by experiencing reduced delays) from the changes that occur in the scheduled demand profile as a result of delays and delay propagation.

Keywords: airport delays, network of airports, delay propagation

INTRODUCTION AND BACKGROUND

Until the early 1990s, major air traffic delays were largely confined to a relatively small number of airports. However, the strong growth in the number of airport operations that took place during that decade and, again, between 2003 and 2007, has led to system-wide congestion problems. By 2007, the worst year for delays in aviation history, demand levels at many airports in the United States and, to a lesser extent, in Europe, were close to or exceeded the capacities of these airports for several hours each day, especially during the peak summer season. As a result, not only did delays at individual airports reach record levels, particularly on days when less-than-ideal weather conditions prevailed, but congestion also spread readily on such days, propagating throughout airline fleets and affecting large parts of the airport system. In fact, at least one airline has computed "multipliers" [Beatty 1998] for delays incurred during different parts of the day. In this study it is claimed, for instance, that one hour of delay suffered by an aircraft early in the day may result in seven hours of delay for the airline's entire fleet, as that initial delay propagates to other aircraft and airports later in the day through late flight arrivals and late connections. As is well-known, the net cost of congestion in this tightly inter-connected and over-scheduled network of airports and aircraft is enormous. Estimates for the United States for 2007 range from $12 billion (Air Transport Association 2009) to $41 billion (U.S. Congress Joint Economic Committee 2008) – with the latter estimate purporting to include both the direct costs of the delays to the airlines and their passengers and the indirect and induced costs that these delays cause to the airline industry (e.g., by forcing the industry to increase the scheduled gate-to-gate time of flights) and to other industries.

The study of how delays propagate within the airport system and of the (occasionally counter-intuitive) queuing phenomena associated with such propagation is the primary focus of the model and results presented in this paper. Airline schedules include some slack, both in the planned gate-to-gate time-lengths of flights and in the "turnaround times" on the ground between consecutive flights of any given aircraft. For instance, while an aircraft may be scheduled to be on the ground between flights at some particular airport for 45 minutes, it may require only 35 minutes to turn-around, thus providing a slack of 10 minutes. But these slacks are typically insufficient to absorb the longer delays that typically occur on a daily basis, thus leading to the propagation of delay. For example, if an aircraft is delayed by one hour in departing from airport A, it will almost certainly be late in arriving at its next airport B; the late arrival at B may also result in a late subsequent departure of that aircraft from B – leading to the dreaded announcement of a delay "due to a late-arriving aircraft".

The Approximate Network Delays model (AND) described here is a stochastic and dynamic queuing model of a network of airports designed to compute approximately delays at each of the individual airports in the network and, more important, how these delays propagate from one airport to another over the course of a day or other desired time period. It treats the airports in the network as a set of interconnected individual queuing systems. Delays at any specific airport may impact delays at other airports in the network, as aircraft execute their daily flight schedules (or "itineraries") by flying from one airport to another.

AND employs a combination of a numerical queuing model (its "queuing engine", QE) and a delay-propagation algorithm (DPA). The QE computes delays at individual airports and the DPA tracks the propagation of these delays and their impact on subsequent airline

operations at all the other airports in the network. Because of its analytical queuing engine, AND does not require multiple runs, as simulations do, to generate statistically reliable performance metrics. Moreover, because it is fast, the model can be used, to explore at a macroscopic, approximate level a large number of scenarios, alternatives and policies. In addition to questions about the existing system, one can study, for example, the impact of hypothetical developments such as, in increasing degree of difficulty: a new runway that boosts capacity at a hub airport; a future air traffic management system (e.g, NextGen in the United States) that increases capacity at all airports; the imposition of "slot limits" at certain key airports, limiting the number of aircraft movements that can be scheduled per hour; and the initiation of a congestion pricing scheme at a selected sub-set of airports.

From the technical point of view, problems involving network-wide airport congestion are difficult to analyze. Steady-state queuing models are inapplicable for all but the crudest approximations because airport demand typically varies strongly with time-of-day (as does airport capacity occasionally) and the dynamic characteristics of airport queues are of central interest. Deterministic flow models with bottlenecks fail to capture another essential characteristic, namely the stochastic aspects of airport operations. Continuous models with stochastic elements (e.g., heavy-traffic, diffusion approximations of queues at individual airports) are also of limited usefulness at the network level, especially if one is interested in tracking the delays experienced by individual aircraft executing their daily itineraries.

It is therefore not surprising that most of the (few) available models of queuing in airport networks are simulations. The National Airspace System Performance Analysis Capability (NASPAC) of MITRE CAASD (Frolow and Sinnott 1989) was one of the first NAS simulation models used by the FAA. Wieland (1997) describes the Detailed Policy Assessment Tool (DPAT), a low-level-of-detail simulation of the national airport network also developed at MITRE CAASD as a successor to NASPAC. Given a set of daily capacity and demand profiles at each airport in the system, DPAT estimates delays for any flight between any two airports. However, DPAT lacks input information on aircraft itineraries, i.e., the sequence of airports that each aircraft will visit on a given day and thus does not capture the impact of delays at any one airport on delays in the rest of the system. In addition, the queuing model, which is used to model airport congestion, is very simple and approximate.

Far more microscopic simulation models of air traffic operations at the national level have also been developed. The state-of-the-art models in this category are ACES, the Airspace Concept Evaluation System (Raytheon 2003) and FACET, the Future ATM Concepts Evaluation Tool (Bilimoria et al 2000) both used currently by NASA and the FAA. These are agent-based simulation tools that model with high detail the entire National Airspace System (NAS), but require a great amount of computational time, as well as extensive input preparation. For example, a run of ACES with a full-scale representation of the NAS typically requires several hours – and it takes multiple runs to obtain statistically reliable estimates of the parameters of interest. These models are therefore ill-suited for the types of policy-oriented, macroscopic issues that AND is designed to address.

Turning to analytical models, Peterson, Bertsimas and Odoni (1995a, 1995b) have used a combination of a deterministic fluid-flow approach and a semi-Markovian model of airport capacities to model non-stationary queuing networks configured as a hub-and-spoke network with a single hub airport at its centre. Like AND, their model accounts for changes in demand rates caused by earlier delays, but is limited by the fact that it considers only a single hub.

Long et al (1999) developed a national-scale airport network model, LMINET, which has some features similar to those of AND, including the modelling of individual airports as $M(t)/E_k(t)/1$ queuing systems. However, the approximation approach they used to compute numerically the queuing statistics is completely different from the one used in AND and obtains estimates of only the expected value and standard deviation of delay at each airport. More important, LMINET does not use information on aircraft itineraries and therefore does not capture the propagation of delays through the network by tracking individual aircraft as they cover their daily routes. Finally, Tandale et al (2009) recently described a family of queuing network models of the NAS designed to facilitate the study of the effectiveness of 4-dimensional trajectory-based operations in reducing traffic delays. The model does not consider the propagation of delays through the disruption of scheduled aircraft itineraries. This last paper also includes a good review of other simulation and queuing models of airport and airspace networks developed for various applications different from those of the AND model.

The contributions of our paper are of both a quantitative and qualitative nature. Quantitatively, it describes a very fast approximate analytical approach to the study of national-scale or international-scale airport queuing networks that keeps track of delays to individual aircraft and flights, as well as computes the “downstream” effects, temporally and geographically, of congestion at one or more airports in the network at any time during a day of operations. The model is also of more general interest, in that it presents an iterative “decomposition approach” that greatly reduces computational effort and thus may make feasible the study of some other types of large queuing networks. Qualitatively, the model and its results provide insights on the complex interactions that take place when the airport system operates under the strain of widespread congestion. These interactions may have quite surprising effects and may benefit certain types of flights or airlines, while penalizing others. Some of the observations detailed in the example of Section 3 of this paper are made for the first time, to the best of our knowledge.

The outline of the remainder of the paper is as follows. Section 2 provides a general description of the model, outlines its overall logic and presents in some detail the Queuing Engine and the Delay Propagation Algorithm. Section 3 describes the implementation of the model for a large group of the busiest airports in the United States and describes a comparison of three scenarios that illustrates the types of analyses that the model can be used for and some of the insights that it provides. Finally, Section 6 provides a summary of conclusions and of ongoing or planned work with the model.

MODEL DESCRIPTION

AND is a model that was originally developed conceptually and implemented in prototype form in the PhD dissertation of Malone (1995). It is based on the fundamental observation that airlines fly their aircraft on daily scheduled itineraries that require visits to a sequence of airports. Given the scheduled itineraries of all the commercial aircraft that fly within a regional or national system of airports, it should then be possible to trace the propagation of delays from airport to airport: if a particular aircraft is scheduled to fly from Airport A to Airport B and then to Airport C and departs from A with a long delay, part or all of that delay will be

propagated downstream and result in a late arrival of the aircraft at B and subsequently at C. To operate properly, a model that captures this process should be able to (a) compute delays at each individual airport and (b) trace how delays at individual airports spread through the aircraft itineraries to the other airports in the system. Moreover (a) and (b) should be performed efficiently if the model is to be of practical use. In this section we provide a description of AND, beginning with a brief overview.

Operation of the Model

Let K be the set of airports in the network. We assume the demand rate at the network's nodes is periodic, with period T. Typically, but not necessarily, T is equal to a 24-hour period, beginning at a time when there is little air traffic activity throughout the network (such as 4 a.m. Eastern time in the United States). T is subdivided into m sub-periods, $h_1, h_2, ..., h_m$ of equal length. We have typically used m = 96, i.e., each sub-period is 15-minutes long, but other sub-period lengths may be more appropriate in different application contexts.
The AND model employs a combination of analytical and algorithmic approaches to estimate local delays at nodes in the network and to propagate delays system-wide. The model iterates between its analytical Queuing Engine (QE) and its algorithmic part, the Delay Propagation Algorithm (DPA), as illustrated in Figure 1. The analytical part, QE, employs a stochastic and dynamic queuing model, DELAYS, to compute delays at each airport, as if the airport operated in isolation. However, alternative QEs can be used, if desired. The DPA "propagates" delays from each individual airport to the rest of the network by tracing how individual aircraft are affected by local delays and updating demand profiles at every airport affected by delays taking place at upstream airports on each aircraft's itinerary.

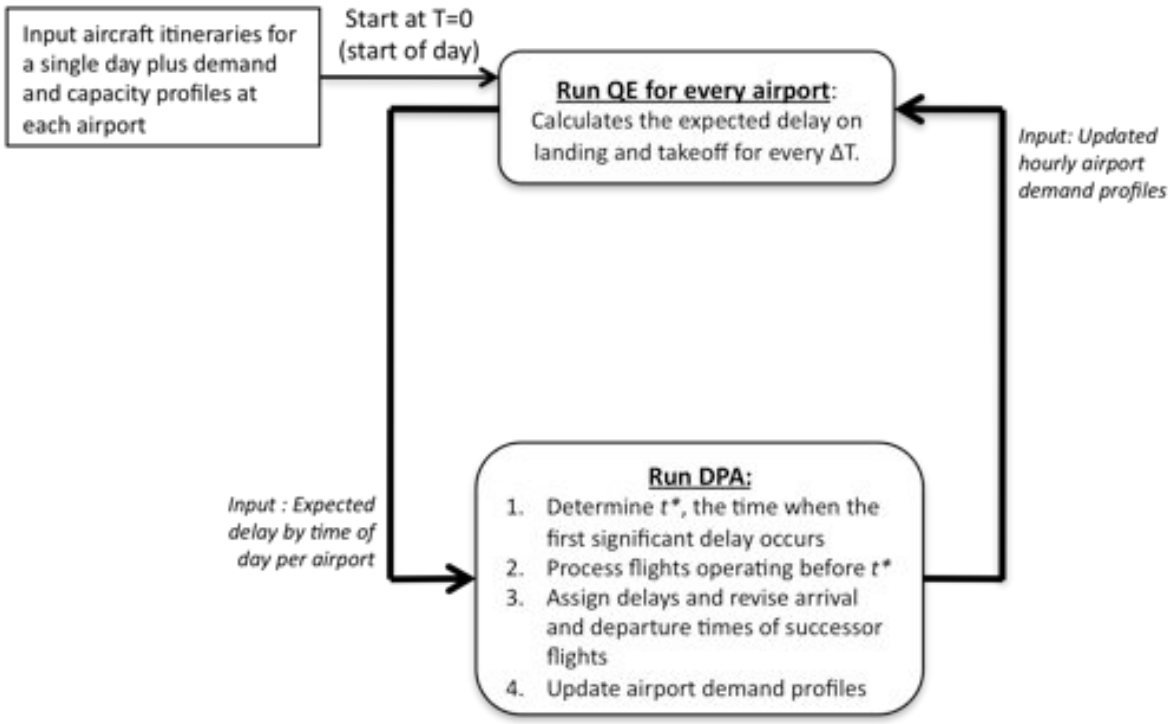

Figure 1 – A schematic representation of the operation of the AND model.

In addition to the QE and DPA, the AND model includes a pre-processor which prepares all of the inputs for the QE and DPA, as shown in Figure 1. The aircraft itineraries contain information about how an aircraft travels through the network of airports over the period. Specifically, it contains scheduled arrival and departure times, the slack time at each airport, and the immediate predecessor and successor flights for each visit to an airport.
Starting at the beginning of the day, the QE is run for every airport in the network separately. Expected delays on takeoff and landing are calculated for every flight. The DPA determines whether delays incurred by any flight are significant enough to warrant propagation, i.e., to affect the original demand rates for any sub-period, h_j, at any airport in the network. If so, the earliest time t^* at which this propagation condition is met is found. All operations (takeoffs or landings) terminating before t^* are unaffected by delays occurring at or after t^*, as are all airport demand profiles for all sub-periods h_j terminating before t^*. These flights are then classified as *processed* by AND and will not be affected by any subsequent iterations of the algorithm. DPA then adjusts the arrival and departure times of all *unprocessed* flights affected by the propagation, as well as the demand rates at affected airports. The QE is then rerun using the updated demand rates and aircraft itineraries. This will, in turn, identify a new earliest time, t^{**}, when the propagation condition is again met and a return to the DPA will be necessary. Note that t^{**} must be greater than t^*, i.e., closer to the end of the time horizon, T. This sequence "QE – DPA – QE – DPA – etc." continues until all delays are eventually propagated through the system and the model reaches the end $t = T$ of its time-horizon.

The Queuing Engine

The AND model treats individual airports as nodes of a queuing network. Arrivals to each queuing system are aircraft requesting service at the corresponding airport, i.e., permission to land or to take off. To avoid confusion, we refer to the rate of requests for landings and takeoffs as the demand rate.

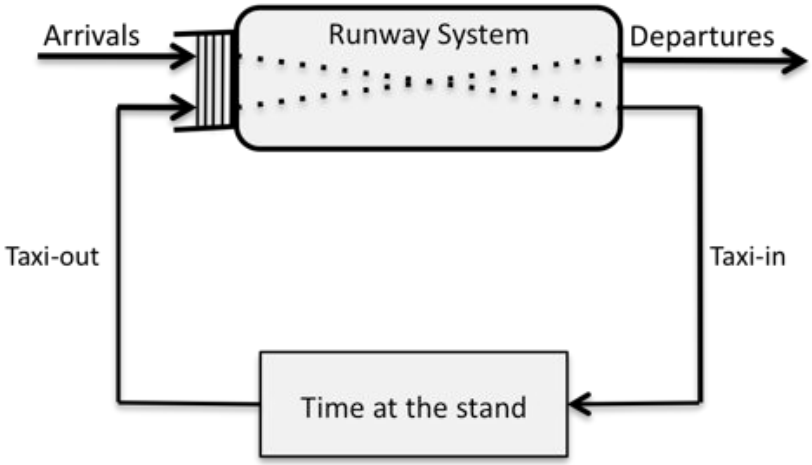

Figure 2 – Conceptual representation of a single airport as a queuing system.

The service at each queuing system is provided by each airport's system of runways. Each runway system, or runway configuration, is modelled as a single server that serves both arrivals and departures. An aircraft needs two services at each airport it visits, one for landing and the other for takeoff, as shown in Figure 2. Demands for access to the runway

system are assumed to be served according to a first-come-first-served (FCFS) discipline. The service rate at each airport/queuing system is equal to the expected number of landings and takeoffs that can take place there per unit of time under continuous demand conditions. This service rate is usually referred to as the *capacity* of the runway system. Infinite waiting line capacity for queuing aircraft is assumed to be available at each airport, so that no demands are ever denied access. Physically, aircraft waiting to take off are queued on the airport's surface (taxiways and, possibly, apron stands), while aircraft waiting to land may be queued in or near the airport's terminal airspace, as well as on the ground at other airports, when a ground delay program for the airport of destination is in effect.

To estimate delays at individual airports, we use a queuing model with a non-stationary Poisson arrival process, time-dependent k^{th}-order Erlang service-time distribution, a single-server, and infinite waiting room, denoted as a $M(t)/E_k(t)/1$ system in queuing theory. Given a demand rate $\lambda(t)$ and a service rate $\mu(t)$, a set of first-order differential equations (often referred to as the "Chapman-Kolmogorov equations") describe the evolution over time of the $M(t)/E_k(t)/1$ queuing system, when the queue capacity is equal to N customers (i.e., aircraft), including the one being served. The typical state-transition diagram of this queuing system consists of $kN+1$ "stages".

In order to approximate an infinite-capacity system, the queue capacity, N, of the system must be sufficiently large so that the probability that the system is full at any time is very small. A practical problem exists, however, in the case of $M(t)/E_k(t)/1$ models for which the number of equations to be solved is $kN+1$: for large k (i.e., when the service times have a small coefficient of variation or, in practical terms, are nearly constant), the number of equations in the system becomes very large, even for modest values of N, and the numerical solution of the system is time-consuming.

For this reason, the Queuing Engine of AND uses an approximation scheme to the $M(t)/E_k(t)/1$ system that solves a set of N+1 difference equations (independent of k), instead of the system of $kN+1$ Chapman-Kolmogorov equations (Kivestu 1976). These difference equations resemble the classical difference equations that describe the evolution of $M(t)/G(t)/1$ systems, except that the "epochs", as defined in the standard approach (Larson and Odoni 2007), are chosen differently. Extensive computational experiments performed by Malone (1995) indicate that Kivestu's approach approximates very accurately the exact solution of $M(t)/E_k(t)/1$ systems and is much faster than the exact (numerical) solution approach, now requiring only a fraction of a second on a typical laptop to compute all the time-dependent delay statistics for a 24-hour period at a busy airport. In addition to approximating accurately the low moments and central moments of the queuing statistics, Kivestu's approach also provides accurate estimates of the state probabilities $P_j(t)$ over time, where $P_j(t)$ is now defined as the probability of having j customers (not stages) in the queuing system at time t. For any given airport a, the expected waiting time in queue, $W_{a,q}(t)$, and expected number of customers in queue, $L_{a,q}(t)$, are approximated through the relationship:

$$W_{a,q}(t) \approx \frac{L_{a,q}(t)}{\mu_a(t)} = \frac{\sum_{j=1}^{N}(j-1)P_{a,j}(t)}{\mu_a(t)} \quad \text{Eq. 1}$$

The Delay Propagation Algorithm

The Delay Propagation Algorithm (DPA) accounts for the interactions among flights by propagating any significant delays that are incurred at an airport to "downstream" flights and airports. Specifically, the DPA performs four functions:

(i) determines if significant delays occur (a delay is "significant" if it will propagate downstream);
(ii) processes connections between flights at each airport;
(iii) adjusts the arrival and departure times of any delayed flights; and
(iv) updates airport hourly demand rates.

In this section, the logic of these four functions is described briefly.

As noted already, we assume that the demand rate at the network's nodes (the airports) is periodic, with period T (usually 24 hours) and vary over the period T. The demand rates are based on the schedule of airport operations (see further below). The expected number of demands and the expected number of service completions when the server is continually busy at airport $a \in K$ in sub-period h_j, $j = 1, 2, ..., m$ are denoted by $\lambda_a(h_j)$ and $\mu_a(h_j)$, respectively. Δh denotes the length of each sub-period.

Because the AND model performs calculations over a finer-grain time scale than Δh, we distinguish between the demand and service rates above for the sub-period h_j and the instantaneous demand and service rates $\lambda_a(t)$ and $\mu_a(t), t \in \Re, 0 \le t \le T$, as used by the *Queuing Engine*. Define $index(t)$ as the function that maps the time of day, t, into the sub-period, h_j, to which it belongs. The relationship between t and h_j, is given by: $j = index(t) = \left\lfloor \frac{t}{\Delta h} \right\rfloor$.

An aircraft's itinerary consists of a set of flights, f, between pairs of consecutive airports. For example, the itinerary {A, B, C, D} consists of a sequence of three flights, A-to-B, B-to-C, and C-to-D. (Note that, according to our definition and without loss of generality, each flight consists of a single "hop".) F is the set of all flights, f, that are scheduled to operate during period *T* in the network. Hence, we define:

$o(f) =$ airport of origin of f
$d(f) =$ airport of destination of f
$SD(f) =$ scheduled departure time of f from $o(f)$
$SA(f) =$ scheduled arrival time of f at $d(f)$
$AD(f) =$ adjusted departure time of f from $o(f)$
$AA(f) =$ adjusted arrival time of f at $d(f)$

Here, by departure and arrival time we refer to the time a flight will request permission to take off or land respectively. This time does not include any delay incurred in the process of

landing or taking off. Thus, $SA(f)$ and $SD(f)$ refer to the times when these requests would have been made according to the original schedule of flight f while the adjusted times, $AA(f)$ and $AD(f)$, refer to the time of the day when f will actually request to land or takeoff. Scheduled and adjusted times may be different due to delays suffered earlier in the day by the aircraft performing flight f. While we set initially $AA(f) = SA(f)$ and $AD(f) = SD(f)$ for all f, some flights may have $AA(f) \neq SA(f)$ and/or $AD(f) \neq SD(f)$ at the end of an AND run, due to the presence of delays in the network of airports. We assume that the immediate predecessor flight, f', to every flight f in every aircraft's itinerary is known, as are the parameters $turn(f)$ and $minturn(f)$. The parameter $turn(f) = SD(f) - SA(f')$, between the arrival of f' and the departure of f, is the scheduled "turnaround time" on the ground of f. Associated with each turnaround time there is also a "minimum turnaround time", $minturn(f)$, which is the smallest amount of time necessary to "handle" (unload, clean, refuel, load, etc.) the aircraft arriving as flight f' and get it ready to depart as flight f. The quantity $slack(f) = turn(f) - minturn(f)$ indicates the "slack" associated with f. The flight f will still depart on time, as long as the arrival of f' is delayed by less than $slack(f)$.

We also define sets G, G_D, and G_A, such that $G \cup G' = G_D \cup G'_D = G_A \cup G'_A = F$. These sets indicate whether any particular flight f has been *processed* or not by AND. (For a definition of "processed", please see further below.) Specifically, $G_D(f) = 1$ if the departure of f has been "processed" by AND, and 0 otherwise; similarly $G_A(f) = 1$ if the arrival of f has been processed; and if both the departure and the arrival of f have been processed, $G(f) = 1$. We call G and G', respectively, the sets of processed and unprocessed flights.

Whenever the DPA is called, e.g., at some time t*, it must determine whether significant delays have occurred among the unprocessed departures and arrivals, G'_D and G'_A. Delay is "significant" if there exists $f \in G'_D$ or $f \in G'_A$ such that f's departure or arrival time must be adjusted. The necessary adjustment to the departure time of f (that is to the time when take-off from an airport is requested) is determined by three quantities:

- $AA(f')$, the adjusted arrival time of the flight f' immediately preceding f in the relevant aircraft's itinerary
- $W_{d(f'),q}(AA(f'))$, the expected delay that flight f' will incur on landing
- $slack(f)$, the slack in the turnaround time between the arrival of f' and the departure of f

Then we have:

$$AD(f) = \max\left[SD(f),\ SD(f) + \left(AA(f') - SA(f')\right) + W_{d(f'),q}(AA(f')) - slack(f)\right] \qquad \text{Eq. 2}$$

Likewise, f's adjusted arrival time depends on:

- $AD(f)$, the adjusted departure time of f
- $W_{o(f),q}(AD(f))$, the delay that f incurs on takeoff
- the time made up in the air flying from $o(f)$ to $d(f)$

Then we have:

$$AA(f) = \max[SA(f),\ AD(f) + W_{o(f),q}(AD(f)) + \left(flight\ time\ \ from\ o(f)\ \ to\ \ d(f)\right) - \left(time\ \ made\ \ up\ \ en\ \ route\ \ from\ o(f)\ \ to\ \ d(f)\right)] \qquad \text{Eq. 3}$$

Given that at least one significant delay occurs, DPA determines when the earliest one of these delays occurs. All flights operating before the earliest significant delay will be unaffected by the delays to be propagated after this delay. Thus, all arrivals and departures occurring before this delay become *processed* arrivals and departures. The time and sub-period at which the earliest delay occurs is found as follows:

$$t^* = \min\left[\min_{f \in G_D'}\left(AD(f)\right), \min_{f \in G_A'}\left(AA(f)\right)\right]$$

$$h^* = index(t^*)$$

To carry out function (iii), DPA propagates delay by making one-step adjustments to the arrival and departure times of those flights that are immediate successors of the flights processed in the current iteration. That is, if flight f' is processed in the current iteration, and f is its immediate successor flight (and has not yet been processed), f's departure time is adjusted using Equation 2. Similarly, if only the departure of flight f' is processed in the current iteration, f''s departure time is adjusted using Equation 3.

If the adjusted arrival time of f falls into a different sub-period h_j than the sub-period into which it was originally scheduled, then the demand rates at $d(f)$ are adjusted, thus performing function (iv) of the DPA. Specifically, let $s = index\left(SA(f)\right)$ and $n = index\left(AA(f)\right)$, i.e., while the arrival of flight f was originally scheduled to take place in sub-period s, this arrival is now expected to take place in sub-period n, with $n \geq s$. Then, if $n > s$, we reduce $\lambda_{d(f)}(h_s)$ by one unit (one fewer expected arrival at the destination airport $d(f)$ of flight f during sub-period s, and increase $\lambda_{d(f)}(h_n)$ by one unit. We perform an entirely analogous adjustment for departures, using $o(f)$, $SD(f)$ and $AD(f)$.

Performance Measures of Interest

Performance measures of interest can be classified as "local" and "network." Local performance measures are those related to delays incurred at each individual airport during a particular visit by an aircraft using that airport, given the (possibly modified) demand rates over the course of the day.

Network performance measures account for the delays that are observed at individual airports, but cannot be attributed to local congestion because they have been incurred earlier in the day at airports previously visited by an aircraft. We also refer to this type of delay as "upstream" or "propagated" delay, while in Europe the term "reactionary" delay is used. AND estimates the upstream delay associated with every arrival and departure at all the airports in the network throughout the day. Given these estimates, more composite performance measures (such as the total amount of upstream delay observed at a specific airport during the course of a day) can be easily computed. AND also routinely computes another network performance measure of practical interest. This is the "fraction of arrivals with expected delay greater than 15 minutes" at each airport and system-wide. This measure is analogous to the well-known statistic published by the US Department of Transportation, which is used in ranking the on-time performance of airlines in the United States. The difference is that the measure computed by AND refers to airport performance, not to airlines. However, airline-

specific on-time statistics can also be computed by AND, if desired, since aircraft itinerary data include airline identification.

We show below how some local and network performance measures are calculated by the AND model. Many more measures can be defined and estimated, if needed, since the model computes all the state probabilities $P_{a,i}(t)$ of having $i\ (=0,1,2,...)$ aircraft in queue at time t for all airports $a \in K$.

Local performance measures include average delay (for specified parts of the day, e.g., for each hour, or for the entire day), maximum expected delay during the day, and the time-averaged number of aircraft waiting to land or takeoff. Examples of network performance measures include:

- Expected upstream delay: a measure of how much of an airport's operating behind schedule can be attributed to congestion incurred earlier in the day in the network. For airport $a \in K$:

$$\sum_{\{f \in F \mid d(f)=a\}} \left[AA(f) - SA(f) \right] \qquad \text{Eq. 4}$$

- Fraction of arrivals with expected delay greater than 15 minutes: this statistic or rating attempts to provide overall "on-time" performance at airports and is patterned after the Department of Transportation's principal measure of on-time performance of airlines – the difference being that the statistic is computed here from an airport's perspective. This fraction is calculated in AND through the expression

$$\frac{\sum_{\{f \in F \mid d(f)=a\}} \delta_a(f)}{\left| \{ f \in F \mid d(f) = a \} \right|} \qquad \text{Eq. 5}$$

where

$$\delta_a(f) = \begin{cases} 1 \ \ if \ \ AA(f) - SA(f) + \text{landing delay of } f > 15 \ \ \text{minutes} \\ 0 \ \ otherwise \end{cases}$$

MODEL IMPLEMENTATION

The AND model has been programmed in java and has been implemented for a network consisting of the 34 continental US airports listed in the FAA's Operational Evolution Plan (OEP) (all airports, excluding Honolulu). Figure 3 shows these 34 airports. The airports and their acronyms are also listed in the Appendix, which also shows the approximate capacities per hour (expected number of landings and takeoffs) of these airports under visual meteorological conditions ("optimum") and under instrument meteorological conditions ("IFR", instrument flight rules). They collectively handled 1.12 billion passengers in 2007 or 73% of the 1.54 billion using airports with scheduled commercial service. A 35th "virtual airport" acts as an un-capacitated source and sink of traffic: any aircraft that operates between an airport in the 34-node network and some other ("external") airport is included in the model by having that aircraft fly to/from the virtual airport. In this way, we account for aircraft that may leave the network at some point during the day and may return to the network at a later time before the end of the day. It takes approximately 100 seconds to run AND with all 34 airports in good weather conditions for one day of operations, 120 seconds with one airport operating with reduced capacity and 140 seconds with six airports operating with reduced capacity.

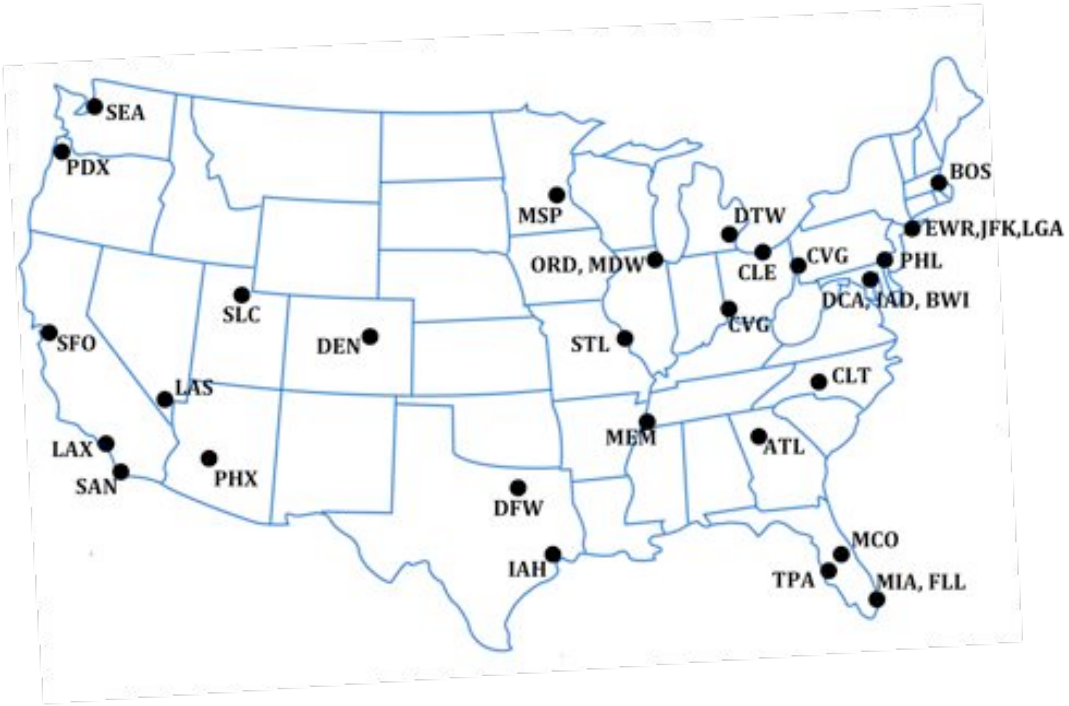

Figure 3 – AND network map.

Inputs

Three types of inputs are critical to the AND model for the period of interest, typically a full day of operations:
(a) *Full daily demand schedules at each airport:* For the 34 US airports implementation of AND, this input was obtained from the FAA's Aviation System Performance Metrics (ASPM) database. The 24-hour period is subdivided into sub-periods of equal length – typically of 15 minutes – and the expected number of demands in each sub-period is indicated. The expected number of demands, in addition to scheduled commercial flights, must also include estimates of the number of general aviation flights expected in each sub-period. These latter estimates can be obtained reliably from historical records for each airport and hour of the day.
(b) *Expected service rates at each airport:* The "service rates" indicate the throughput capacity of each airport's runway system in terms of the expected number of movements that can be processed per sub-period. The service rates can be obtained from analytical or simulation models of airport capacity – see, e.g., Stamatopoulos et al (2004) and Swedish (1981) – or from field-based estimates of airport capacities, as in the FAA's Capacity Benchmark report (FAA 2004). The latter was the source of the service rates used in the detailed example described in the next sub-section. In practice and in the AND model, service rates may vary with time across sub-periods, reflecting potential changes in weather, noise-related restrictions during certain hours of the day, etc.
(c) *Aircraft itineraries*: For every aircraft performing commercial service within the network, this input indicates the detailed sequence of flights to be performed, including scheduled arrival and departure times for the aircraft at each airport. Aircraft itineraries are obtained from the On-time Performance database of the Bureau of Transportation Statistics (BTS) database.

A Detailed Example and Insights into System Performance

We consider next a detailed example based on one of the families of tests conducted with the AND model. The intent is to illustrate the types of inter-airport interactions that can be observed through AND, with emphasis on the insights that the model provides. The particular example described involved the 22 airports listed in Table 2 further below.
Three scenarios will be considered here (Table 1), with each scenario representing different operating conditions at one or several airports. The same day of operations, 8:00am GMT 10/15/2007 to 7:59am GMT 10/16/2007, was used for all three scenarios. In Scenario 1 all airports in the network operate under VFR conditions at optimum capacity levels for the entire day. In Scenario 2 the capacity at ORD is affected by bad weather for the entire day of operations and is reduced to the low IFR level of 36 movements per 15 minutes (see Appendix), while all other airports operate at optimum capacity. In this case, the expected demand at ORD, as viewed at the beginning of the day, exceeds the capacity for most of the day. (Note from the Appendix that ORD, which is already quite congested under optimum operating conditions, faces a large decrease in capacity when moving from optimum to IFR

operating conditions.) Finally, in Scenario 3, six Northeastern airports—BOS, EWR, JFK, LGA, DCA and PHL—are affected by a storm, which decreases the capacity of all six airports to IFR levels (see Appendix) for the entire 24-hour period.

Table 1 – The three scenarios of the example

	Scenario 1	*Scenario 2*	*Scenario 3*
Capacity levels:	All airports in optimum	ORD in low IFR	BOS, DCA, EWR, JFK, LGA, PHL in low IFR

Overall Network Behaviour

When one or more airports operate under low IFR conditions one would expect to observe some increase in upstream delays at all the airports of the network. This is indeed the case when the expected upstream delays estimated by AND for *Scenarios 2* and *3* are compared to those for *Scenario 1*, as shown in Table 2. Similarly, as expected, the fraction of flights that arrive with more than 15 minutes of delay undergoes an often large increase, either because of local congestion at the airport of arrival or due to upstream delay incurred at earlier destinations.

A second observation concerns the large amount of expected upstream delay that ORD experiences in *Scenario 2*, when ORD operates in low IFR, while all other airports operate at their highest capacity. While this may appear counter-intuitive at first, it is perfectly logical when one considers the planned aircraft itineraries of the two airlines, American and United, which use ORD as a hub. Many of the aircraft of these two airlines have daily itineraries of the type "…ORD – other airport(s) – ORD – other airport(s) – ORD …", i.e., involve repeated visits to ORD during the course of the day. This means that, if an aircraft suffers a significant delay on one or more of its visits to ORD (e.g., as a result of poor weather, as in the case of *Scenario 2*), this delay will affect the entire subsequent schedule of that aircraft and will propagate through part or all of the day. In fact, the more times the aircraft lands at and takes off from ORD under adverse conditions, the more upstream delay is generated for subsequent visits to ORD. Essentially, the local delay suffered at ORD on each visit to the airport "feeds on itself" and becomes upstream delay, as far as the remaining scheduled flight legs for the day are concerned. In a nutshell, ORD causes high upstream delays onto ORD itself!

This phenomenon is much less intense at non-hub airports, as suggested by Table 3, since the number of aircraft that visit non-hubs, like LGA, several times during the course of a day is much smaller than in the case of hubs, like ORD. In fact, as can be seen from Table 3, although the fraction of arrivals that suffer a delay of more than 15 minutes is approximately the same for ORD and LGA under *Scenarios 2* and *3*, respectively, the expected upstream delay in ORD is 17 times greater, while there are only twice as many flights operating through ORD.

Table 2 – Expected upstream delay for the 24-hour period and fraction of arrivals with more than 15 minutes of delay for a set of 22 out of 34 airports of the network for Scenarios 1-3.

	Scenario 1		*Scenario 2*		*Scenario 3*	
	Upstream Delay (min)	*% of Arrivals with >15min delay*	*Upstream Delay (min)*	*% of Arrivals with >15min delay*	*Upstream Delay (min)*	*% of Arrivals with >15min delay*
ATL	99	0.2%	2089	2.2%	2173	3.2%
BOS	233	1.9%	2208	8.1%	3200	14.8%
CLE	75	0.9%	1457	6.5%	1165	10.2%
CLT	25	0.0%	1534	5.7%	1322	5.9%
DCA	127	0.8%	1868	8.6%	2388	17.5%
DEN	53	0.2%	3815	8.4%	633	2.0%
DFW	37	0.1%	1843	3.0%	1011	2.6%
DTW	71	0.2%	1677	4.1%	1186	4.7%
EWR	4	0.0%	1578	4.5%	3311	71.7%
IAH	18	0.0%	1181	3.1%	935	3.9%
JFK	281	29.8%	1194	30.6%	1674	62.6%
LAS	96	0.8%	1557	5.1%	934	2.7%
LAX	231	1.2%	2519	5.3%	1625	3.1%
LGA	0	0.0%	2602	8.6%	1653	84.5%
MCO	120	1.1%	779	3.1%	1151	6.6%
MIA	81	1.2%	829	8.7%	781	7.5%
MSP	37	0.3%	2393	8.1%	555	2.8%
ORD	157	0.5%	28522	87.3%	2486	4.2%
PHL	1	0.0%	1444	6.1%	276	0.7%
PHX	84	0.5%	1245	2.9%	821	1.7%
SEA	87	1.3%	1270	6.1%	669	2.6%
SFO	211	1.5%	2043	7.4%	1591	4.5%

Local versus upstream delays

Figure 4 shows how the total delay suffered by all aircraft during the day breaks down between local and upstream delay at 9 of the 22 airports in the network under each of the three different scenarios. It is clear that when all airports operate under optimum conditions (Scenario 1) the local delay at each airport accounts for most of the delay at each airport. When ORD operates in low IFR (Scenario 2), the upstream delay at all other airports becomes much more significant, as it ranges from 10% (for JFK) up to 88% (for BOS) of the total delay at each airport. Similarly, under Scenario 3, the upstream delays at every airport account for a much greater fraction of the total delays, compared to Scenario 1. A further interesting observation is that the upstream delay under Scenario 3 is higher than under Scenario 2 for only 5 out of the 9 airports shown, although there are six airports operating in low IFR under Scenario 3 and only one (ORD) under Scenario 2. This underscores the central role that ORD played in 2007, as a major and often-congested hub, in generating delays and “spreading” these delays throughout the national air transportation system.

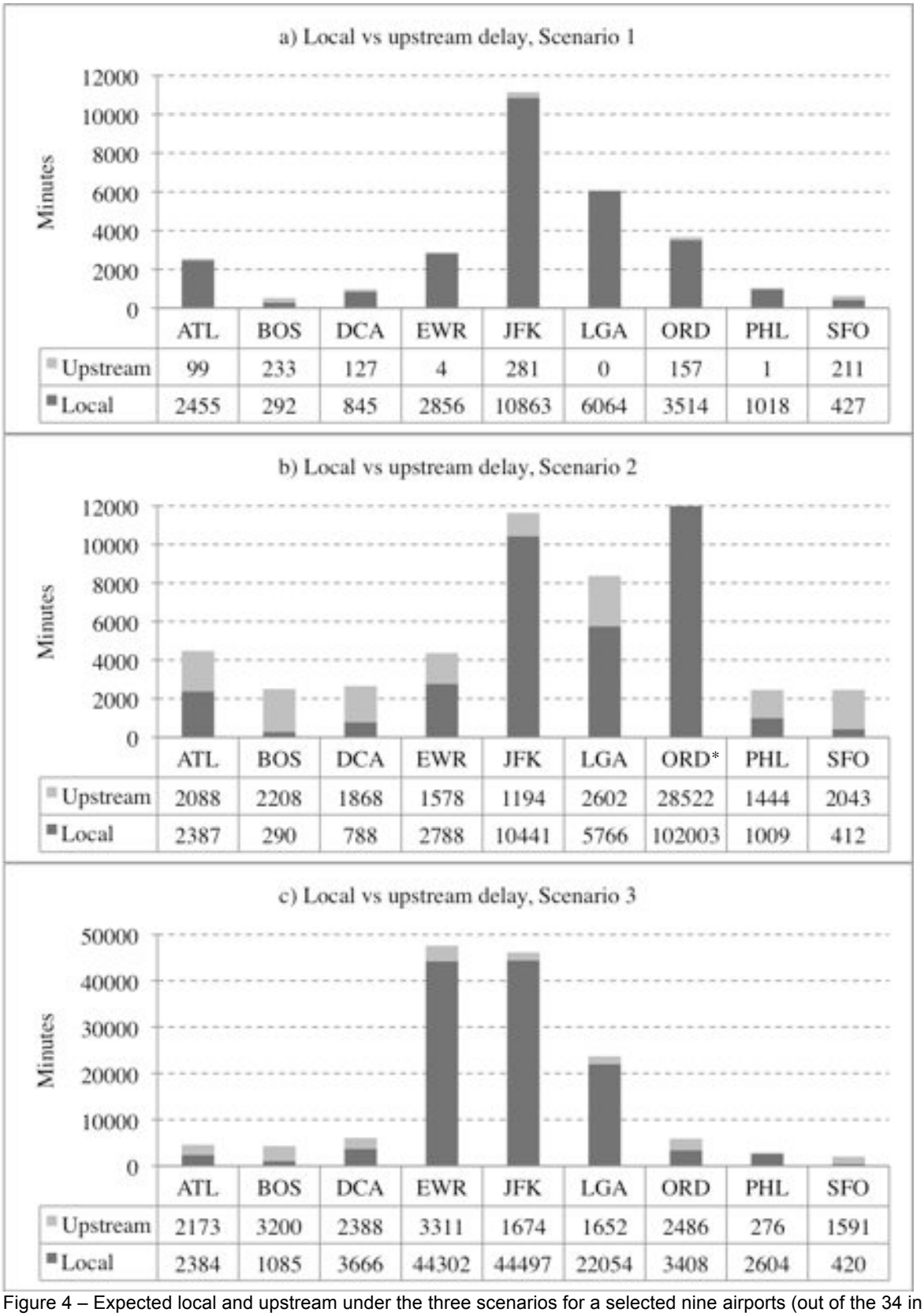

Figure 4 – Expected local and upstream under the three scenarios for a selected nine airports (out of the 34 in the network). *Note: In Figure b the bar for ORD is not shown to its full extent; the bar extends to approximately 130,000 minutes.

Shift in demand profiles and smoothing of the expected local delay

In order to perform a more detailed and airport-specific analysis and to understand better the effects of the propagation of delays on individual airports, we shall now concentrate on just two congested airports, ORD and LGA – a hub and a non-hub, respectively. For Scenario 2, Figure 5a shows (i) the total number of landings and of take-offs that are initially scheduled in every 15-minute interval of a day of operations at ORD ("scheduled", in blue) and (ii) how this demand profile changes as a result of the propagation of delays ("adjusted", in red). Please note that the "adjusted" and the "scheduled" demand profiles coincide during the early part of the day, until roughly 10:30. One can observe that, after the morning hours, the adjusted demand shifts very noticeably to the right by comparison to the scheduled demand. This is a consequence of the same phenomenon noted earlier: when congestion occurs at ORD, aircraft that must fly through the airport several times a day, suffer delays earlier in the day, which propagate to their subsequent "visits" to ORD, i.e., generate "upstream" delays for these aircraft. This moves the instants when they demand access to ORD to later times than in the original schedule. The net effect is to spread demand more evenly than originally scheduled. The propagation of delays thus effectively creates a "smoother" demand profile at ORD later in the day – and especially towards the end of the day – when, as can be seen in Figure 5a – most of the peaks are "evened out" and the demand level approximately "tracks" the available capacity.

In an analogous way, Figure 6a shows the changes in the demand profile at LGA under Scenario 3. Although many of the peaks of the scheduled profile are again evened out, the smoothing effect in this case is not nearly as strong as that observed at ORD. As previously explained, the reason for this is that ORD, in contrast to LGA, is a hub: there are many more aircraft which fly through ORD more than once during a 24-hour day than through LGA.

The shift in the demand profile of ORD under Scenario 2 – and resulting smoothing and spreading of demand over more hours of the day – leads to a much lower level of expected local delay and a much more even distribution of profile of delay by time-of-day, as shown in Figure 5b. Note that the estimated local delay, based on the originally scheduled demand profile at ORD peaks at almost 160 minutes of expected delay for a movement scheduled at 21:45. But, with the "adjusted" demand profile shown in Figure 5a, the local delay peaks at approximately only 60 minutes: a flight arriving at ORD around 21:30 on that particular day would be expected to experience a local delay of about 60 minutes upon arrival, as opposed to about 2.5 hours with the original demand profile. The explanation is that delays earlier in the day have caused many flights originally scheduled to arrive or depart around 21;30 to request service later in the day, thus making the time interval around 21:30 less congested than originally. The analogous effect at LGA (Figure 6b) is again not as strong as at ORD, because LGA is not a hub and fewer aircraft visit it more than once in a typical day.

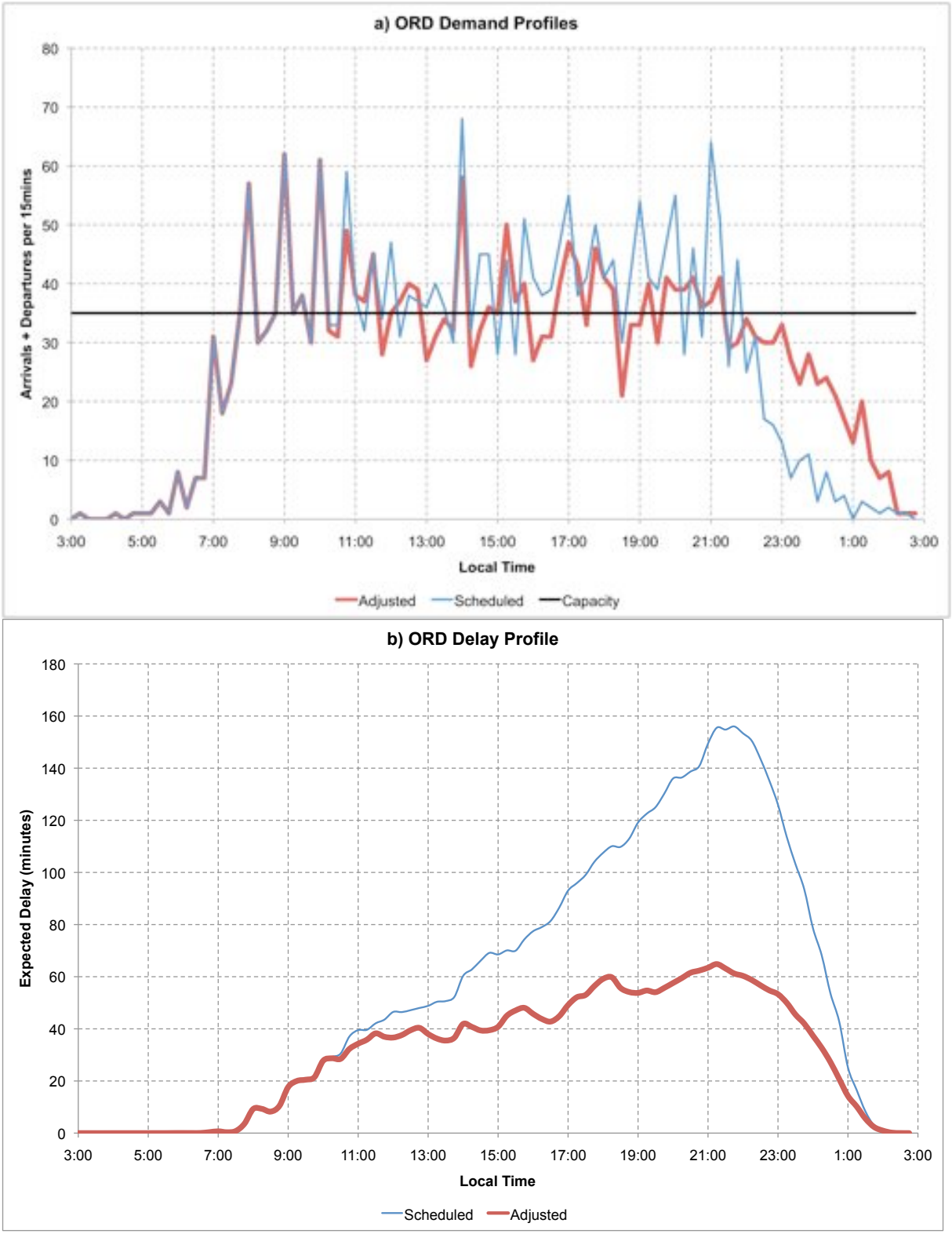

Figure 5 – a) Demand profiles and b) the corresponding expected local delay at ORD under Scenario 2.

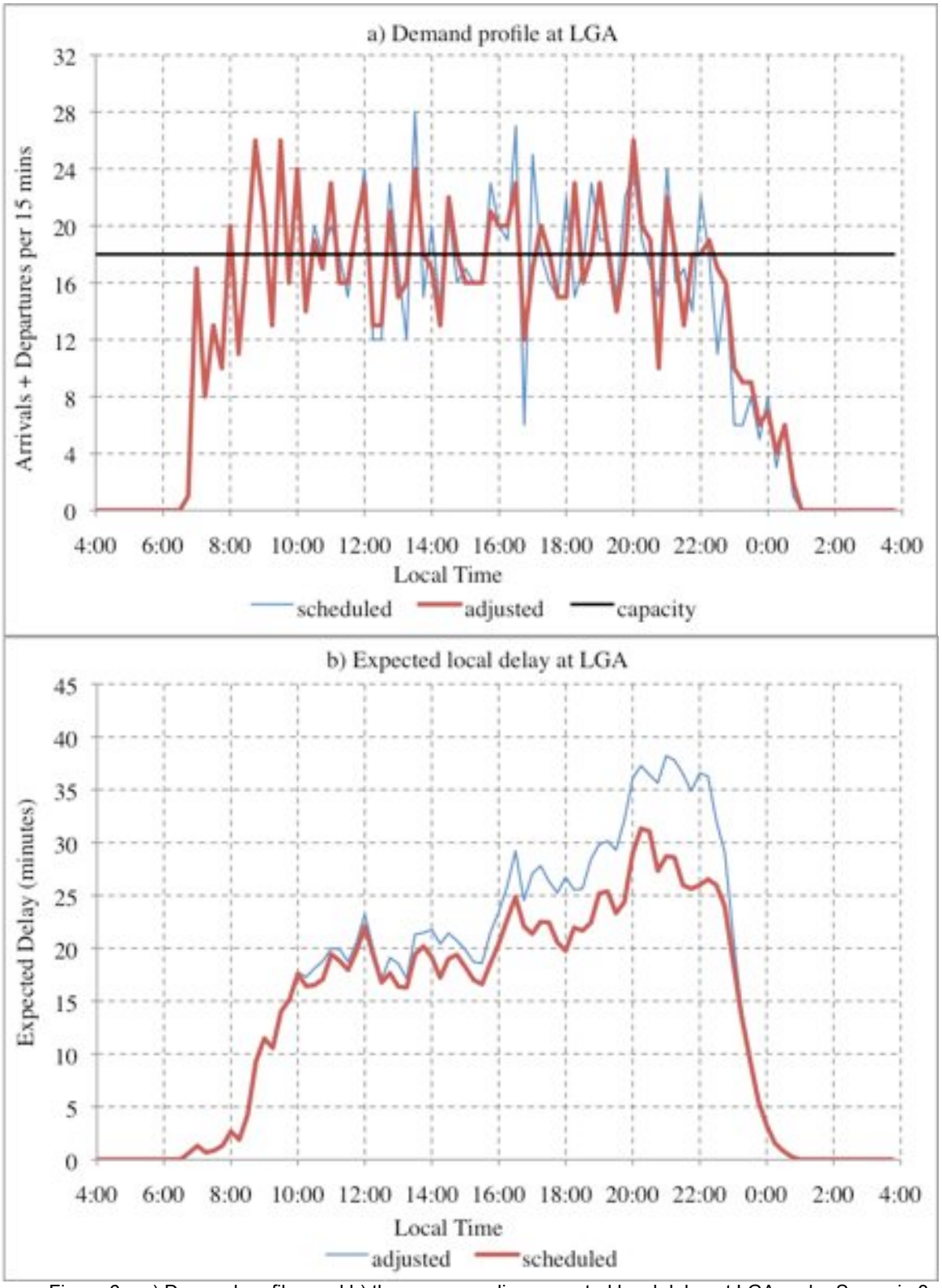

Figure 6 – a) Demand profiles and b) the corresponding expected local delay at LGA under Scenario 3.

Total Delay

Figure 7a shows the expected total arrival delay suffered by every flight that operates in ORD under *Scenario 2* during the day examined. Note that this expected total delay is the sum of the expected upstream delay that each aircraft experienced at locations that it previously visited during the day and of the expected local delay suffered upon arrival at ORD. Clearly, as also shown in Figure 7b, the expected total delay increases as the day progresses and as aircraft experience delays at various points in the network that can not be absorbed by the slack in their scheduled itineraries. Moreover, we observe in Figure 7a that, not only does the expected total delay increase on average later in the day (until about 22:00 hours), but its variability increases greatly as well. This is also shown in Figure 7b by the increase of the standard deviation of the total delay. Hence, as the day progresses, the expected deviation from schedule, i.e. the expected total delay, increases, while the reliability of the schedule deteriorates – i.e., the uncertainty about flight arrival and departure times increases. This is a phenomenon familiar to experienced US air travellers.

The graph labelled “expected local delay based on the adjusted schedule” in Figure 7a is the same as the one showing the “adjusted” expected local delay in Figure 5b. Note that in Figure 7a the points corresponding to many flights fall on top of this graph, i.e., the total delay experienced by many flights is equal to the “expected delay based on the adjusted schedule”. These are flights performed by aircraft that operate for the *first time* during the day at ORD. Thus, their expected total arrival delay consists only of the expected local delay upon arrival at ORD. It is evident that these flights “benefit” from the smoothing of the demand profile at ORD. They experience smaller delays than would be expected under the original schedule in ORD. This underscores the usefulness of models like AND: had we not accounted for delay propagation and the resulting shift of the demand profile in ORD, we would have estimated that flights by aircraft operating for the first time in ORD would suffer an expected delay equal to that shown by the *“expected local delay based on the initial schedule”* line of Figure 7a.

Figure 7a also shows the breakdown of expected total delay for those flights performed by aircraft not visiting ORD for the first time. For each of the points lying above the *“expected local delay based on the adjusted schedule”*, the corresponding flight suffers a local expected delay equal to the value of the *“expected local delay based on the adjusted schedule”* plus an upstream delay equal to the difference between its total expected delay, as shown in Figure 7a, and its expected local delay.

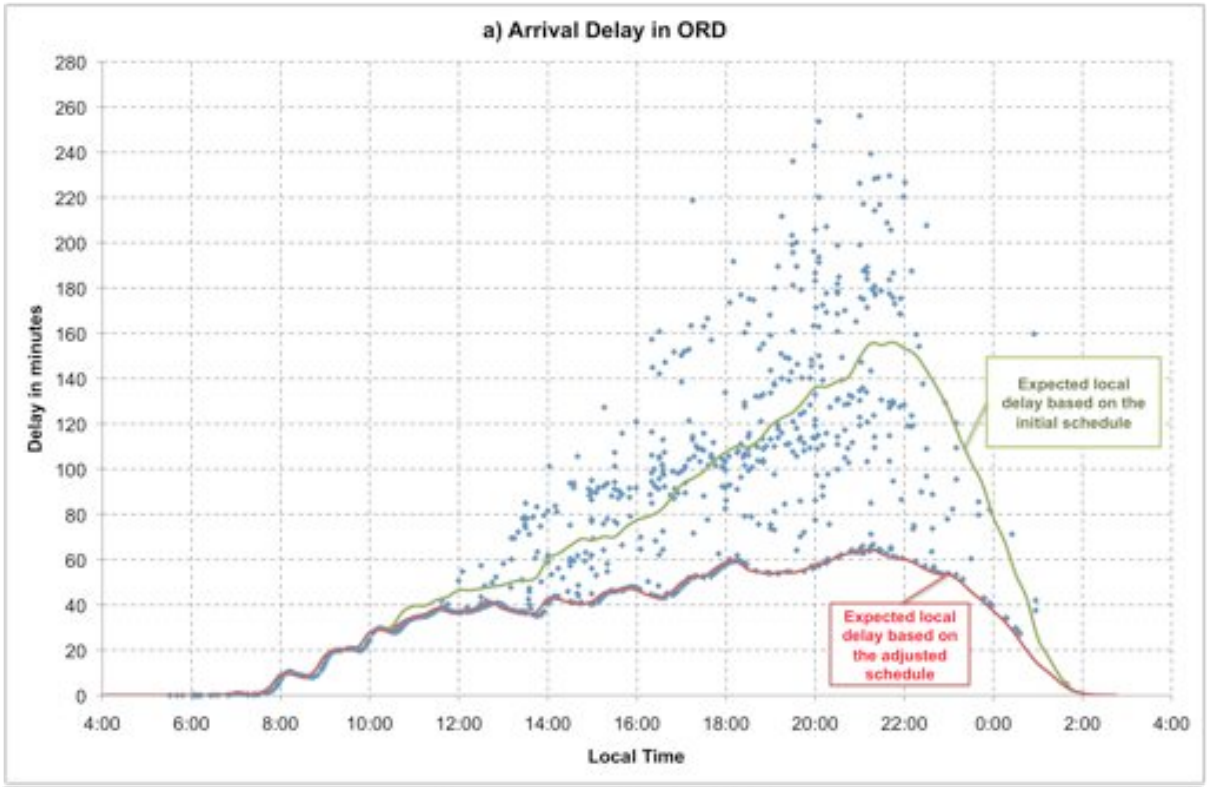

Figure 7 – a) Expected total arrival delay per flight and b) average arrival delay and standard deviation per hour at ORD under Scenario 2.

CONCLUSIONS AND FURTHER WORK

We have presented an approach for modelling the system-wide effects of congestion in a large network of major airports, taking into consideration the stochastic nature and strongly time-varying characteristics of airport demand and capacity. Because of the efficiencies gained through the decomposition approach adopted and the speed of its queuing engine, the AND model can be a powerful tool for exploring from a macroscopic perspective the implications of a broad range of alternative policies and strategies concerning demand management or capacity expansion interventions in a regional or national system of airports.

- The model is best used to study relative performance of such a system under numerous scenarios and hypotheses.

Ongoing work involves further calibration and validation of the model. For example, the model's results are sensitive to the setting of the "slacks" in ground turnaround times. We are therefore conducting extensive data analysis on scheduled and minimum turnaround times for different types of aircraft, airlines, and airports (hubs and non-hubs). AND's results have also been checked extensively for logical consistency. For instance, we have verified through an analysis of ASPM data from Newark and New York JFK airports (a) the powerful shift of demand toward the later hours of the day shown in Figure 5a for ORD on days with poor weather and (b) the increase of the average total delay as the day progresses as shown in Figure 7. These analyses will be reported in detail in a subsequent paper.

It should be noted, however, that the critical test for AND is whether it can re-produce trends and behaviours that are observed in practice in the NAS system, not whether it can reproduce exactly the actual delays measured. The reason is that AND does not model airline reactions to congestion and does not capture other sources of delays such as aircraft mechanical issues, en route congestion and propagation of delays due to crew scheduling constraints. Another critical test is whether AND produces results that are of the same order of magnitude as delays measured in practice. Again, ASPM data analyses performed to date are reassuring on this point.

Two important extensions of the model have also been implemented. One is the development of an alternative queuing engine, a deterministic D(t)/D(t)/1 model [Hansen et al 2009]. Running AND with this engine provides a useful lower bound on expected delays for cases where congestion is high and the deterministic model generates meaningful delay estimates. The second is the modelling of Air Traffic Flow Management Ground Delay Programs (GDPs) on days when congestion is particularly severe. Again these enhancements will be reported in a future paper.

Examples of AND applications planned for the future may include, in increasing order of difficulty, studies of the impacts of:

- Changes in airport capacity resulting from local interventions (e.g., a new runway at a specific airport)
- Effectiveness of system-wide programs (e.g., NextGen) designed to increase capacity or to increase predictability (e.g., 4-D trajectories) in the national airport and ATM system.
- Efforts to alleviate the impact of upstream congestion on delays at "downstream" airports through air traffic flow management actions.

- New configurations of airline networks, such as configurations that may de-emphasize hubbing or the use of certain highly congested airports.
- Airline strategies that reduce (or increase) the amount of time aircraft are scheduled to spend on the ground at airports between arrival and departure, or the required minimum time for "turning around" aircraft on the ground, or the amount of flight schedule "padding", i.e., the amount by which nominal flight times are increased to account for expected delays and to increase schedule reliability.
- Airport demand management policies, such as slot limitations and congestion pricing.

Finally, we have recently received from EUROCONTROL the data necessary to implement and apply AND to a network consisting of the 34 busiest airports in Europe. This will make it possible to compare the characteristics of the US and the European airport systems with regard to the magnitude of delays at individual airports and the extent to which local delays propagate throughout the respective networks.

APPENDIX

The hourly service rates in the table below are based on the estimates provided in the FAA Benchmark Report (2004). "Optimum" capacity is achieved in visual meteorological conditions (VMC) while "IFR" indicates the airport capacity in instrument meteorological conditions (IMC).

Table 3 – List of US airports included in the AND network

		Hourly Service Rates	
		Optimum	**IFR**
Atlanta Hartsfield	ATL	237	202
Boston Logan	BOS	123	93
Baltimore	BWI	106	71
Cleveland	CLE	80	64
Charlotte	CLT	130	110
Cincinnati	CVG	120	110
Washington Reagan	DCA	72	70
Denver	DEN	210	162
Dallas Fort Worth	DFW	270	193
Detroit	DTW	184	145
New York Newark	EWR	84	66
Fort Lauderdale	FLL	60	56
Washington Dulles	IAD	135	113
Houston	IAH	120	112
New York JFK	JFK	75	67
Las Vegas	LAS	102	70
Los Angeles	LAX	137	124
New York LaGuardia	LGA	78	74
Orlando	MCO	144	117
Chicago Midway	MDW	64	64
Memphis	MEM	148	132
Miami	MIA	116	96
Minneapolis St Paul	MSP	114	114
Chicago O'Hare	ORD	190	144
Portland	PDX	116	80
Philadelphia	PHL	104	96
Phoenix	PHX	128	118
Pittsburgh	PIT	152	150
San Diego	SAN	56	50
Seattle	SEA	80	60
San Francisco	SFO	105	72
Salt Lake City	SLC	130	113
St Louis	STL	104	70
Tampa	TPA	102	75

REFERENCES

Air Transportation Association. (2009). Cost of ATC Delays. Retrieved April 2009, from Air Transportation Association: http://www.airlines.org/economics/cost+of+delays.

Beatty, R., R. Hsu, L. Berry and J. Rome (1998). Preliminary Evaluation of Flight Delay Propagation through an Airline Schedule, Proceedings of the 2nd USA/Europe Air Traffic Management R &D Seminar, Orlando, Fl.

Bilimaria, K.D., B. Sridhar, G.B. Chatterji, G. Sheth and S. Grabbe (2000). FACET: Future ATM Concepts Evaluation Tool, Proceedings of the 3rd USA/Europe Air traffic Management Seminar, Naples, Italy.

Federal Aviation Administration (2004). Airport Capacity Benchmark Report, Washington, DC.

Frolow, I. and J. Sinnott (1989). National Airspace System Demand and Capacity Modeling, Proceedings of the IEEE, 77 (11), 1618 – 1624.

Hansen, M., T. Nikoleris, D. Lovell, K. Vlachou and A. Odoni (2009). Use of Queuing Models to Estimate Delay Savings from 4D Trajectory Precision, Proceedings of the 8th USA/Europe Air Traffic Management Research and Development Seminar (ATM2009), Napa Valley.

Kivestu, P. (1976). Alternative Methods of Investigating the Time-Dependent M/G/k Queue, S.M. Thesis, Dept. of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA.

Larson, R.C. and A. Odoni (2007). Urban Operations Research, Dynamic Ideas, Belmont, MA.

Long, D., D. Lee, J. Johnson, E. Gaier and P. Kostiuk (1999). Modeling Air Traffic Management Technologies with a Queuing Network Model of the National Airspace System, Technical Report NASA/CR-1999-208988, NASA Langley Research Center, Hampton, VA.

Malone, K. (1995). Dynamic Queuing Systems: Behavior and Approximations for Individual Queues and Networks, Ph.D. Thesis, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA.

Peterson, M.D., D.J. Bertsimas and A.R. Odoni (1995a). Models and Algorithms for Queueing Congestion at Airports, Management Science, 41, 1279-1295.

Peterson, M.D., D.J. Bertsimas and A.R. Odoni (1995b). Decomposition Algorithms for Analyzing Transient Phenomena in Multiclass Queueing Networks in Air Transportation, Operations Research, 43, 995-1011.

Raytheon ATMSDI Team (2003). Airspace Concept Evaluation System: Build 2 Software User Manual, NASA Ames Research Center, Moffett Field, CA.

Stamatopoulos, M.A., K.G. Zografos and A. R. Odoni (2004). A Decision Support System for Airport Strategic Planning, Transportation Research C, 12, 91-118.

Swedish, W.J. (1981). Upgraded FAA Airfield Capacity Model, Vols. 1 & 2, Technical Report MTR-81W16, The MITRE Corporation, McLean, VA.

Tandale, M.D., P. Sengupta and P.K. Mennon (2009). Queuing Network Models of the National Airspace System, Working Paper (submitted for publication), Optimal Synthesis Inc., Los Altos, CA.

Wieland, F. (1997). Limits to Growth: Results from the Detailed Policy Assessment Tool, AIAA/IEEE Digital Avionics Systems Conference – Proceedings, 2, 9.2.1 – 9.2.8.

U.S. Congress Joint Economic Committee. (2008). Your Flight Has Been Delayed Again: Flight Delays Cost Passengers, Airlines, and the U.S. Economy Billions, Washington, DC.