The Static Stochastic Ground Holding Problem

Michael O. Ball [*] Robert L. Hoffman [†]
Amedeo R. Odoni [‡] Ryan M. Rifkin [§]

1 Introduction

It is well-known that demand for runway operations (landings and take-offs) often exceeds available capacity for periods of several hours at major airports throughout the world. This is primarily true under unfavorable weather conditions (poor visibility or storms of various kinds) but it happens occasionally under good weather conditions as well, especially in the United States, where some of the airports are "overscheduled" due to the hubbing practices of major airlines or due to seasonal increases in traffic.

When the expected demand for landings is predicted to exceed the arrival capacity of a major airport in the United States for a significant period of time, the Federal Aviation Administration (FAA) uses various air traffic flow management (ATFM) measures to smooth out the arrival flow and bring arrival demand in line with capacity. Ground holding is the most important of these methods. The idea is simple: it is preferable to have a flight wait on the ground at its point of origin rather than to have it circle the airport at its destination, unable to land. Therefore, if it is known with certainty, or at least with high probability, that a flight will be unable to land due to lack of capacity, it may be advantageous to hold the flight on the ground at its point of origin. Ground holding saves fuel costs and preserves safety margins by relieving airborne congestion.

The FAA adpoted ground holding as a commonly-employed strategy in the early 1980's. For each possibly capacitated airport, the FAA generates an estimate, or forecast, of capac-

[*]University of Maryland
[†]University of Maryland
[‡]Massachusetts Institute of Technology
[§]Massachusetts Institute of Technology

ity for the day. Treating this forecast as deterministic, the FAA assigns ground delays to incoming aircraft so that the arrival flow will match the forecasted capacity. If the forecast is accurate, this ensures that air holds (planes forced to wait in the air at their destination due to lack of landing capacity) will be kept to a minimum. This policy will be referred to as the *deterministic* ground holding policy.

An airport acceptance rate (AAR) is the number of flights that can be landed at a given airport in a given unit of time (strictly speaking, this is a capacity not a rate, nonetheless, it is the established terminology in ATFM). The short-term forecast of capacity on which ground holding policies are based is known as the forecast of airport acceptance rates (AAR) and is usually given for each hour over several hours. For instance, it might be predicted that over a six-hour period, the AAR of an airport will be 36 flights for the first hour, then 30 fights for each hour thereafter, perhaps reflecting worsening weather conditions.

Forecasted weather conditions for an airport are converted to runway configurations and subsequent AARs with reasonable accuracy. The main problem with the deterministic ground holding policy is that it ignores the highly stochastic nature of the weather conditions that ultimately determine the AARs. For instance, if the forecasted AARs turn out to be lower than the AARs that acutally materialize (perhaps due to improved weather conditions), then (in retrospect) too much ground holding has been applied and valuable airport capacity goes unutilized. Similarly, if the forecasted AARs are higher than the actual AARs, then demand will exceed capacity and there will be airborn holding that (again, in retrospect) could have been replaced with ground holding.

A large body of work exists on deterministic versions of the ground-holding problem. See e.g., Terrab [**?**], Vranas [**?**], Vranas [**?**], Bertsimas and Vranas [**?**], Vranas, Bertsimas and Odoni [**?**], and Bertsimas and Stock [**?**] (this list is by no means exhaustive).

In this article, we look at a *static stochastic* model for the single-airport ground holding problem: stochastic, in that it explicitly takes into account the stochastic nature of future capacity, and static, in that it requires all decisions over a given time horizon to be made in advance. The seminal paper on the static stochastic ground holding problem is “Solving Optimally the Static Ground-Holding Policy Problem in Air Traffic Control,” by Richetta and Odoni [**?**]. An integer program is developed that deals with stochasticity by assuming the existence of a probabilistic distribution of "scenarios", or possible realizations of capacity. By treating arrivals as flows rather than as individual flights and by making use of the fact that the linear programming relaxation always yields an integer solution, the model is solvable for reasonable problem sizes.

The present paper, based on Hoffman [?] and Rifkin [?], makes three main contributions. The first is to introduce a new model for the static stochastic ground holding problem. This model produces the same solutions as the Richetta model but takes advantage of pre-processing and post-processing to reduce the number of decision variables by an order of magnitude. The second contribution is a proof that the integer program associated with the model is dual network, which implies that the model can be solved to optimality in polynomial time via network methods and that the LP relaxation yields an integer solution.

The third contribution of this paper is that the adopted perspective is entirely consistent with the recently-initiated (January 1998) Collaborative Decision-Making (CDM) approach to assigning ground holds to individual flights (see [?]). Prior work on ground delay problems focussed on assigning delays to individual flights in order to optimize a system-wide objective function. But it is difficult to justify the optimizaton of a single objective across all airlines: in reality, the air traffic system consists of a large number of individual users (airlines) with diverse, often conflicting, objectives.

Under CDM, the FAA and the airlines have jointly taken the alternate view that the FAA is responsible for forecasting airport capacity and, in the event that a GDP is warranted, paritioning arrival resources into "arrival slots" to be distributed amongst inbound flights in an equitable manner. For instance, if the AAR for the first hour of a GDP is 30 fights, then an arrival slot is created for every two seconds of that hour. In essence, fairness (as defined by CDM) is achieved by awarding the earliest slots to the flights with the earliest scheduled arrival times (as scheduled by the Official Airline Guide, created weeks earlier). For example, suppose that the capacity of an airport whose normal capacity is 50 flights per hour is forecasted to drop to 36 flights for the first hour and 30 flights for the next two hours. Then the first 36 flights on the list of scheduled flights would be assigned to the first 36 flights, the next 30 flights on the list would be assigned to the next 30 slots, and so on. Each airline is then given an opportunity to redistribute its flights among the slots it has been awarded, subject to certain flight eligibilities. (In practice, there are many practical considerations such as flights that cannot make their appointed slot time and flights exempted from FAA-assigned ground delay. See [Rif: add ref] Hoffman, Ball, Hall, Odoni, Wambsganss.)

This is precisely the view in this paper, as well: the output of the stochastic ground delay model we will describe is the *number* of arrival slots that should be made available during each time interval of a GDP. The specific assignment of arrival slots and subsequent ground delays to flights is accomplished through CDM.

In Section 2, we formulate the problem more precisely, state our assumptions, and present our model. In Section 3, we prove that the integer program defined by the model is dual network. In Section 4, we present an example of the model's behavior. Finally, in Section 5, we draw conclusions and point out future research directions.

2 Model Formulation

The static stochastic ground holding problem assumes that the only element of uncertainty is the arrival capacity at the airport in question; demand and travel times are deterministic and known in advance. We assume that the time interval of interest consists of $T \in \mathcal{Z}_+$ time periods. For each time period, there is a demand D_t, which is the number of flights predicted to arrive at the airport in time interval t, if there were no capacity restrictions. We also assume that probabilistic information about the uncertain capacity is available in the form of Q *scenarios*, M_q, for $1 \leq q \leq Q$, where $M_{q,t}$, $1 \leq tT$, is the arrival capacity of the airport during time t, if scenario q were realized. We assume that the probability of the q'th scenario occurring, p_q, is known. Let $c_g > 0$ be the cost of ground holding a single plane for one time period and let $c_a > 0$ be the cost of one period of air delay for a single plane. We assume that these costs are available, linear in the length of the delay, and that $c_a > c_g$ (if not, there is no need for ground holding). To ensure feasibility, we add a $T + 1$'st time period during which the airport has an arbitrarily large capacity (delays may extend beyond the T'th interval - indicating that the original time interval was not sufficient to ensure a recovery from lost capacity).

Our decision variables are A_t, $1 \le t \le T+1$, the number of planes that should be schedule to arrive at the airport's airspace at time t. We assume that planes arriving at the terminal airspace are landed in a First-Come First-Served (FCFS) manner. We also introduce auxiliary variables G_t, and $W_{q,t}$, for $1 \le t \le T+1$, $1 \le q \le Q$, where G_t is the number of flights that wish to land at time t but are (ground) delayed at their point of origin by a ground delay, and $W_{q,t}$ is the number of flights held in the air from time period t to $t+1$ by an airborn delay.

Within the model, the use of the G_t variables is equivalent to assuming that all travel times are instantaneous and that planes are held on the ground until immediately prior to their arrival time. [This issue needs to be addressed in more detail - better model development would help.] Because travel times are assumed deterministic and known, the actual departure times can be easily recovered by subtracting the travel time from the scheduled arrival time.

Our objective is to minimize the expected value of the sum of the air and ground delay costs. This gives rise to the following integer programming problem:

$$\min \sum_{t=1}^{T} c_g G_t + \sum_{q=1}^{Q} \sum_{t=1}^{T} c_a p_q W_{q,t} \qquad (1)$$

$$A_t - G_{t-1} + G_t = D_t \qquad t = 1, \ldots, T+1$$

$$(G_0 = G_{T+1} = 0) \qquad (2)$$

$$-W_{q,t-1} + W_{q,t} - A_t \ge -M_{q,t} \qquad t = 1, \ldots, T+1$$

$$q = 1, \ldots, Q$$

$$(W_{q,0} = W_{q,T+1} = 0) \qquad (3)$$

$$A_t \in \mathcal{Z}_+, W_{q,t} \in \mathcal{Z}_+, G_t \in \mathcal{Z}_+ \qquad (4)$$

The objective function (1) is the sum of the (fixed) ground delay costs and the (expected) air delay costs. Constraint set (2) defines conservation of flow for planes departing from their destination airport: all planes that were delayed from the previous time period (G_{t-1}), plus new demand for this time period (D_t), must either be delayed until the next time period (G_t) or allowed to arrive (A_t). Constraint set (3) performs a similar function for the air delay queue at the capacitated airport. The "$\geq$" rather than an "$=$" is necessary because there may not be enough planes available to fill up the airport's capacity.

We note that if $G_T = 0$ or, for any q, $W_{q,T} = 0$, then the assumption of sufficient recovery capacity does not hold and an extension of the original time interval of the GDP (beyond T) should be considered.

This model is closely related to Richetta's model (see [**?**]), but is much simpler. In Richetta's model, the decision variables were of the form $X_{i,j}$, $1 \leq i \leq j \leq t$, the number of planes that were originally scheduled to arrive at time i that were rescheduled to arrive at time j. Additionally, he used slightly superlinear ground holding costs, in order to avoid solutions where some planes were ground held excessively and unfairly. If one insists that planes be allowed to arrive and land in the order in which they were originally scheduled (FCFS discipline), the required ground delays for individual flights can be easily recovered via post-processing. This corresponds the FAA's policy of fairness.

We emphasize that with simple, linear-time post-processing, this model produces solutions identical to Richetta's model but is substantially faster to solve. Additionally, this model admits a formal proof that the LP relaxation of the IP is guaranteed to yield integral solutions (see Section 3), and is consistent with current CDM initiatives.

We now turn to the cost parameters c_a and c_g. An inspection of the model indicates that the optimal solution depends only on the *ratio* of these parameters $r = \frac{c_a}{c_g}$. A possible objection to this model is the necessity to specify this ratio. We emphasize that this ratio need not represent actual relative dollar costs of ground and air delay — indeed, the relative costs of air and ground delay may vary across airlines and even across flights by the same airline. Instead, the cost ratio r need only be a quantification of the FAA's willingness to trade ground delay for air delay, taking into account the tradeoff between its desire to serve the industry efficiently and its operational and safety concerns.

3 Theoretical Results

In this section, we show that the static stochastic ground holding problem can be solved in polynomial time. In particular, we show that the integer program defined in the previous section is a dual network flow problem. As immediate corollaries, the constraint matrix associated with this IP is totally unimodular, and the LP relaxation yields integral solutions.

We say that an m row, n column $(0, 1, -1)$ matrix M is a *network matrix* if there exists a directed tree R on $m+1$ nodes and a one-to-one mapping of the rows of our matrix onto the edges of R with the property that each column of the matrix corresponds to the characteristic vector of a path in R. This definition was introduced in [**?**] and is explored in [**?**]. The more familiar *node-arc incidence matrices*, with a single 1 and -1 in each column, are a special case of this construction. Network matrices are desireable because they are totally unimodular and give rise to integral polyhedra.

Theorem 1 *Let A^T be the transpose of the constraint matrix associated with the static stochastic ground holding problem. Then A^T is a network matrix.*

Proof Sketch:We will illustrate the construction of the tree R for the case $T = 2, Q = 2)$ (two time periods and two scenarios) from which it will be clear how to extend the construction to an arbitrary number of time periods and scenarios. The general construction is explicitly formulated as a matrix transformation in [**?**], thus providing a more formal proof.

For a two-time period, two-scenario problem with slack variables, the primal problem is:

$$
\begin{array}{llllllllllllllll}
\min & & & & \multicolumn{6}{l}{c_g \cdot G_1 + c_g \cdot G_2 + c_a \cdot W_{11} + c_a \cdot W_{12} + c_a \cdot W_{21} + c_a \cdot W_{22}} & & & & & & \\
 & A_1 & & & +G_1 & & & & & & & & & & = & D_1 \\
 & & A_2 & & -G_1 & +G_2 & & & & & & & & & = & D_2 \\
 & & & A_3 & & -G_2 & & & & & & & & & = & D_3 \\
 & -A_1 & & & & & +W_{11} & & & & -S_{11} & & & & = & -M_{11} \\
 & & -A_2 & & & & -W_{11} & +W_{12} & & & & -S_{12} & & & = & -M_{12} \\
 & & & -A_3 & & & & -W_{12} & & & & & & & = & -M \\
 & -A_1 & & & & & & & +W_{21} & & & & -S_{21} & & = & -M_{21} \\
 & & -A_2 & & & & & & -W_{21} & W_{22} & & & & -S_{22} & = & -M_{22} \\
 & & & -A_3 & & & & & & -W_{22} & & & & & = & -M
\end{array}
$$

Where M is any sufficiently large constant (e.g., $M = \sum_{t=1}^{T} D_t$). We write the matrix A associated with this problem as below;

$$
A = \begin{bmatrix}
A_1 & & & G_1 & & & & & & & & & & & \\
 & A_2 & & -G_1 & G_2 & & & & & & & & & & \\
 & & A_3 & & -G_2 & & & & & & & & & & \\
- - A_1 & & & & & W_{11} & & & & -S_{11} & & & & & \\
 & -A_2 & & & & -W_{11} & W_{12} & & & & -S_{12} & & & & \\
 & & -A_3 & & & & -W_{12} & & & & & -S_{13} & & & \\
- - A_1 & & & & & & & W_{21} & & & & & -S_{21} & & \\
 & -A_2 & & & & & & -W_{21} & W_{22} & & & & & -S_{22} & \\
 & & -A_3 & & & & & & -W_{22} & & & & & & -S_{23}
\end{bmatrix}
$$

Note: actually, A^T is a $(0, 1, -1)$ matrix but, for ease of association of each column with its respective variable, we have substituted the corresponding variable name for each 0, 1,

-1 entry in its column. To show that A^T is a network matrix, we must find a directed tree R on 16 nodes such that every column of A^T (equivalently, every *row* of A) corresponds to the characteristic vector of a path in R. Figure 1 shows the required tree. We have labeled each arc of R with the corresponding primal variable. It can be verified that every row of A corresponds to the characteristic vector of a path in R. For example, the fifth row of A corresponds to the path formed by traversing arc S_{12} backward, then arc W_{12} forward, then arcs A_2 and W_{11} backward. Figure 2 shows the graph G spanned by the tree R. We label each arc with the associated dual cost coefficient (in the primal, the associated RHS element). This construction extends to an arbitrary of number of scenarios and time periods. □

Corollary 2 *The constraint matrix associated with the static stochastic ground holding problem is totally unimodular, its LP relaxation yields an integral solution, and the stochastic ground holding problem can be solved in polynomial time.*

The dual problem can be recast in a more familiar form as a min-cost flow problem by negating the objective function and multiplying the dual LP in equality form by N, where N is the node-arc incidence matrix associated with R with a single row deleted. Figure 3 shows the resulting flow problem for the previous example ($T = 2, Q = 2$). The labeled arcs can accommodate either positive or negative flow, as they are associated with equality constraints in the original primal problem; the flow must be nonnegative on the unlabeled arcs, which are associated with dual slack variables. The unlabeled arcs are zero-cost arcs and the unlabeled nodes are neither sources nor sinks.

We note that the primal problem cannot be recast as a network flow problem. This can be verified by applying standard network recognition algorithms (see [**?**] or [**?**]). See [**?**] for further discussion of this issue. In [**?**], it is shown that scenario-stochastic min-cost flow problems in outerplanar graphs can always be recast as dual min-cost flow problems.

Figure 1: The dual tree R

Figure 2: The dual tree R, with columns of A^T

Figure 3: The transformed min-cost flow problem

Test Case	Numb flights	T = Numb periods	Min per period t	Time (sec)	Simplex Iteration	Node B&B
1a	624	12	60	0.07	51	0
1b	624	48	15	0.30	267	0
2a	624	12	60	0.07	59	0
2b	624	48	15	0.37	289	0
3a	624	12	60	0.05	55	0
3b	624	48	15	0.20	223	0

Table 1: SGH model performance

4 Experiments

In [**?**], Richetta and Odoni performed computational experiments in which they compared the quality of solutions generated by their stochastic model to those of the deterministic algorithm, and to the passive algorithm of no ground delays, under various cost and scenario assumptions. The primary conclusion was that in many cases, the stochastic algorithm found solutions with only slightly more ground delay than the passive algorithm, with much lower total expected costs. Since our model finds identical solutions, their results apply.

To demonstrate the tractibility of our model, we used CLPEX 4.0 with default settings on a Sparc Station 10 to solve three realistic instances of the stochastic ground holding problem, each comprised of a demand profile and three AAR scenarios with varying capacities (in the range of 30-60 flights per hour). Since each data set spanned a 12-hour time interval and 624 flights, these problem instances represent unusually long GDPs at a major airport. Each data set was solved in two ways: once with the time horizon divided into 60-minute time periods and once with 15-minute time periods, for a total of six test cases (1A, 1B, 2A, 2B, 3A, 3B). In each test case, we set $c_g = 2.0$ and $c_a = 5.0$.

The results in Table 1 show that the problem is highly tractable. The integer solution was obtained in zero nodes of the mixed integer program algorithm of CPLEX, empirically confirming the integer solution can be obtained directly from the linear program relaxation. The largest number of iterations of the simplex procedure was 289 and the longest run time was barely more than half a second. Note that the run time is almost linear in the coarseness (length of) the time periods.

Next, we explore the solution of an additional test case to gain a qualitative understanding of the solutions generated by the model.

In this example, we know that a storm is approaching but we are confronted with three different forecasts as to its severity and timing. Figure 4 shows the three corresponding (possible) realizations of capacity, along with the arrival demand over the hours in question. Scenario 1, with probability 0.4, is the most severe prediction. It predicts the early arrival of a medium length storm, with an aftershock at period 7. Scenario 2, with probability 0.3, predicts a moderate storm, while Scenario 3, also with probability 0.3, predicts a mild storm of short duration.

We explore several values of the cost ratio $R = \frac{c_a}{c_g}$, the ratio of air delay cost to ground delay cost. Figure 5 shows, for each value of R, the optimal schedule, and the distribution of both ground and (expected) airborne delays. Note first that as R increases, the optimal schedule grows progressively more conservative, allowing fewer and fewer flights to arrive during the earlier, possibly congested periods. The ground and air delay profiles further demonstrate this theme: as R increases, arriving planes have experienced progressively more ground delays, and expect to receive progressively fewer airborne delays. Note that at $R = 25$, no airborne delay can occur. $R = 25$ therefore corresponds to a maximally conservative schedule; increasing R still further will have no additional effect. If the FAA were to apply the deterministic policy to a forecast which consisted of the lowest capacity of any scenario at each time period, this is the schedule that would be produced. Declaring airborne delays to be 25 times as costly as ground delays is almost certainly excessive in practice; the $R = 25$ schedule is shown to indicate how far R must be increased before no airborne delays can occur (at $R = 20$, there are still some expected airborne delays). The high value of R needed to eliminate all airborne delays indicates that applying the deterministic policy to pessimistic forecasts roughly corresponds to optimizing expected costs under an unrealistically high ratio of air delay cost to ground delay cost, producing excessive ground delays. Figure 6 shows, for each value of R, the distribution of ground and (expected) airborne delays with respect to the original arrival schedule.

5 Conclusion

The stochastic ground holding model presented in this paper finds the optimal trade-off between airborne and ground holding in the formulation of a ground delay program. It represents a substantial simplification of an earlier model developed by Richetta; although it finds the same solutions via post-processing, it uses far fewer decision variables. The fact that the model is dual network allows us to relax the integrality constraints, as we are guaranteed integer optimal solutions directly from the LP relaxation. The combination of these two facts results in a highly tractable model which requires only modest computing power. Additionally, the model is designed to be an integral component of the collaborative decision-making process, and can be easily integrated into existing tools.

Practical implementation of our model requires that the following issues be addressed:

- Exempt flights: In practice, not all flights bound for an airport during a GDP can be assigned a ground delay (e.g., international flights). The effect of these exempt flights is to subtract from the capacity of the airport. Our model can fully accomodate these exemptions by pre-processing the AAR scenarios. See [?] for a formal algorithm.

- Our work here assumes the real$_timegenerationofmultipleweatherscenariosbutdoesnotofferarigorousmethodforquantifyingtheprobabilityofeachscenariooccuring.However,wehavereasontobelievethatthiscanbedonethroughacombinationofspecializedforecastingtechniquesandhistoricalstatisticsonweatherconditionsandtheircorrespondingarrivalcapacitites.$

- Experimentation is required as to the behavior of the model with respect to changes in the cost ratio. An inroad to this has been made in (see [?]).

Finally, an ideal future model for ground delay would incorporate several effects which are not modeled here and are not currently included in ATFM practice, including non-linear delay costs, the interaction between arrival and departure capacity at an airport and the "network effects" (propagation to other airports) of ground holding policies.

Notes to Rif:

(1) figure 1 (dual tree of R) has a $W_{3,1}$ entry that should be $W_{2,2}$

(2) figure 2 (dual tree of R) has a $W_{3,1}$ entry that should be $W_{2,2}$

(3) figure 3: the minus sign in $c_g - c_a$ looks like a superscript for g.

(4) fix formatting of biblio (inconsistent use of caps and italics in titles)

(5) add to biblio: (almost done:) Hoffman, Hall, Ball, Odoni, Wambsganss overview paper on CDM: "Collaborative Decision Making In Air Traffic Flow Management", in preparation. This should be cited along with the original Wambsganss paper as a reference for CDM. (the paper is almost done)

Figure 4: Demand and Scenarios

$R = 1.1$, Ground Delay Shown

$R = 1.1$, Air Delay Shown

$R = 3.5$, Ground Delay Shown

$R = 3.5$, Air Delay Shown

$R = 6$, Ground Delay Shown

$R = 6$, Air Delay Shown

$R = 25$, Ground Delay Shown

$R = 25$, Air Delay Shown

Figure 5: Optimal Arrival Schedules, with Delay Profiles

$R = 1.1$, Ground Delay Shown $R = 1.1$, Air Delay Shown

$R = 3.5$, Ground Delay Shown $R = 3.5$, Air Delay Shown

$R = 6$, Ground Delay Shown $R = 6$, Air Delay Shown

$R = 25$, Ground Delay Shown $R = 25$, Air Delay Shown

Figure 6: The Original Arrival Schedule, with Delay Profiles