THE UNTOLLED PROBLEMS WITH AIRPORT SLOT CONSTRAINTS

Joseph I. Daniel
This paper examines the efficiency and practicality of airport slot constraints using a deterministic bottleneck model of landing and takeoff queues. It adapts this congestion pricing model to determine the optimal timing and quantity of slot permits for any number of slot windows. Aircraft choose their optimal operating times subject to the slot constraints, and airport queues adjust endogenously. The number and length of slot windows affects the congestion levels and efficiency gains. The atomistic bottleneck model is extended to include self-internalizing dominant traffic and atomistic fringe traffic. The model raises questions about the implementation of slot constraints that do not arise in standard congestion models. The theory explains Daniel’s (2009a) empirical findings that slot-constraints at Toronto are ineffective and suggests that recent proposals for slot constraints at US airports would be similarly ineffective. Effective slot constraints require many narrow slot windows, making slot auctions or markets difficult to implement.

**Keywords:** airport congestion, slot constraints, pricing, bottleneck, queuing. (JEL R4, H2, L5, L9)

1. **Introduction**

During the closing months of the Bush Administration, the Department of Transportation (DOT) seriously considered a proposal to impose slot constraints to reduce delays at the nation's busiest airports. Slot constraints are quantity restrictions on the number of landings or takeoffs that airports permit per time period, typically by hour. Nearly all major commercial airports in the US that receive federal funding are currently available on a first-come, first-served basis without requiring prior slot permits.\(^1\) DOT officials apparently believe that slot quantity restrictions are

---

\(^1\) Airports assess landing fees based on aircraft weight, typically between one to five dollars per thousand pounds. These fees have nothing to do with the marginal social cost of serving the aircraft. The social cost consists
simpler to implement than congestion tolls because they think airports need only limit the number of slot permits to the airport’s capacity, then sell, auction, or give the permits away. A slot market is supposed to price and allocate the slot permits efficiently, whereas they believe congestion tolls are too difficult (or politically inconvenient) for airports to assess correctly. Moreover, slot permits are supposed to avoid the problem of imposing different toll schedules on dominant airlines that already internalize their self-imposed delays than on fringe airlines that behave atomistically.

Previously unnoticed problems with airport slot constraints call into question the effectiveness and practicality of many actual and proposed implementations of slot-constraint systems as a means of congestion management. These problems do not arise in standard congestion models that underlie most existing policy analysis. The standard models implicitly assume that traffic rates are constant within each time period (slot window) and that congestion does not carry over from one period to another. Under these assumptions, the common practice of issuing slot permits with hourly slot windows in quantities equal to hourly capacity makes some sense. If traffic shifts significantly within slot windows and queues develop in one slot window that spill over into the next, however, then it is desirable to use a model that captures these features of the problem.

In this paper, I develop a bottleneck model with multiple slot windows in which the airport authority chooses the timing of slot windows and the quantity of landings or takeoffs (operations) to permit during each window. Aircraft choose when to operate within the slot windows to minimize the costs of their schedule and queuing delays. A structural model of congestion has state-contingent queues that evolve endogenously in response to traffic adjustments. The congestion externality causes the traffic rate to exceed the service rate during a portion of each slot window, even when the number of slot permits is within the airport’s capacity for each slot window. Airports can allow or prevent traffic from operating but they cannot precisely regulate the traffic rate within slot windows unless they limit the duration of each slot window to a single service interval, which is probably impractical. Slot constraints work by repeatedly turning on the traffic flow, letting the queue fill to a certain level, shutting off

primarily of the delays aircraft impose externally on other aircraft. Weight-based landing fees encourage too many operations by smaller aircraft compared to marginal-cost congestion tolls that promote efficient levels of operations by internalizing the external delays.
the traffic flow, and letting the queue clear. As the airport authority is able to add more slot windows, it constrains aircraft to more narrow operating windows and the queue clears more frequently. This limits the accumulation of aircraft in the queue and reduces queuing delay. The model also addresses the optimal timing of dominant and fringe airline operations, the efficient allocation of slots among dominant and fringe airlines, and the effect of slot constraints on the efficiency and distribution of surpluses between dominant airlines, atomistic fringe aircraft, and the airport authority. A policy section discusses the problems with current and proposed implementations of slot constraints, and the practical issues involved in designing and implementing an efficient slot system. The paper focuses on the more basic issues of timing and quantity of slot permits rather than how to design auctions or markets to distribute the permits.

A brief preview of five propositions I derive from the model is as follows: 1) An effective slot-constraint system requires numerous narrow slot windows that force traffic to spread out over the peak period. 2) A slot-constraint system that holds the quantity of slot permits to the airport capacity over a single slot widow covering the entire peak period is completely ineffective. 3) In an unconstrained equilibrium, dominant airlines schedule a fraction of their aircraft to operate at the service rate in the periods just before and after the peak. These aircraft fully internalize their delays, while the remaining dominant aircraft behave atomistically. The fraction of internalizing aircraft varies from one to zero as fringe demand elasticity varies from zero to negative infinity. 4) If the airport authority has complete control of the allocation of slots, it will separate the dominant and fringe operations to enable the dominant airline to fully internalize. If it is unable to enforce this separation, then the dominant airline will schedule some aircraft atomistically, depending on the elasticity of fringe demand. 5) The first-best optimum requires one slot permit and window for every service interval.

2. Background

The recent proposal under consideration at DOT provides for auctioning slots permits that would be valid for a period of ten years. Each year one tenth of an airport's slots would expire and be re-auctioned. Airlines could resell slots, and presumably markets would develop for trading slots. There are various proposals concerning the type of auctions to use to initially distribute the slots. Since the values of slots depend on their combination with other slots, these auctions must be able to value an enormous number of potential slot combinations. These auctions would potentially be the largest and most complex ever conducted. The fundamental
justification for slot auctions is that they are supposedly simpler than congestion tolling because an auction or market determines the price of slots rather than an administrative agency. Issuing many undifferentiated slot permits with wide slot windows reduces the dimensionality of potential slot combinations, thus making combinatorial slot auctions or slot markets more feasible. Since airlines are free to choose when to operate within the slot windows, however, wider windows with more undifferentiated permits reduce the airport’s control over the timing of traffic and lengths of queues. There is a fundamental tradeoff between feasibility of slot auctions or markets and reduction of congestion. DOT and other proponents of slot-constraints largely ignore this issue because standard congestion models implicitly assume that traffic rates are constant within slot windows. Consequently, hourly slot windows appear to be adequate.

Most of the world’s airports, but not those in the US, are governed by the International Air Transit Association’s (IATA) Worldwide Slot Guidelines (WSG). Airports that are classified as highly congested have slot coordinators who issue and distribute slot authorizations on a semi-annual basis. According to the guidelines, the fundamental considerations in allocating slots are preserving historical patterns of use, preventing “confiscation” of incumbent airlines’ claims on slots, and allocating slots to new entrants only from new airport capacity. Twice a year, airline representatives submit slot requests that substantiate their past slot usage, and airport slot coordinators make preliminary slot distributions, then airline and airport representatives meet at three-day slot conferences to trade and finalize slot allocations. Slots may be traded in one-for-one exchanges, but may not be sold. The airline industry appears to have designed this slot system as a means of restricting entry. DOT intends that its proposal will preserve freedom of entry, but the system is still more restrictive than congestion tolling under a first-come, first-served rule.

Any optimal system of congestion management must address two fundamental issues: setting the traffic quantity so that the marginal social cost of operations equals their marginal social benefit; and controlling the timing of traffic to minimize its total social cost. The standard congestion models and most real-world slot constraint systems attempt to address the first issue but not the second. The slot permits in the DOT proposal, the previous slot-constraint systems in the US, and the systems following the WSG all grant authority to operate during specific time windows—usually one hour intervals. These wide time windows provide flexibility to accommodate random variation in operating times, facilitate exchange of slots, and reduce
administrative and compliance costs, but they also make it impossible to achieve optimal traffic patterns and can render slot constraints largely ineffective. For example, Daniel (2009a) shows that Toronto's Pearson International Airport experiences significant congestion in spite of a slot-constraint system that follows the WSG.

The way airports collect and present traffic data supports the standard models’ erroneous assumption of constant traffic rates within slot windows. Government and consulting reports typically aggregate airport traffic rates by hour. This practice hides the rapid fluctuations in traffic rates over shorter periods of time that are responsible for a significant amount of airport delays. Airports do not routinely collect data on actual time spent in landing or takeoff queues. Instead, the primary measures of aircraft delays are on-time arrival and departure statistics. These report the number of aircraft that are more than fifteen minutes late relative to the aircraft’s scheduled operating time at the gate. Aside from the obvious issue of not recording delays of less than fifteen minutes, on-time operating statistics do not reflect the additional time that airlines add to their scheduled travel times to allow for queuing delays on landing or takeoff, for traffic jams on the tarmac while trying to access the gates, or any of the other many regular causes of aircraft delay. Atlanta and Chicago O’Hare, for example, average approximately twenty minutes of queuing delay per aircraft, which is much higher than other US airports, while their on-time arrival statistics are no worse than most other major airports. The hourly data paints a misleading picture that is consistent with the standard model’s gradual fluctuations in airport traffic rates, while minute-by-minute data on traffic and delay patterns show rapid fluctuations that are inconsistent with the standard model, but are consistent with the bottleneck model.²

3. Review of the literature

The standard congestion pricing model specifies an average (private) travel cost function for traversing a highway segment that is an increasing function of the current traffic volume during some fixed time period. The function is analogous to the average cost function of a firm that produces trips along the highway segment. The marginal (social) cost function lies above the average cost function and is also increasing in traffic volume. The vertical difference between average and marginal cost is the additional travel time that one more vehicle imposes externally

² Daniel and Harback (2009a) document the rapid fluctuation of traffic patterns using minute-by-minute data on traffic and delays at the twenty seven largest US airports. Daniel and Pahwa (2000) compare the performance of the standard model, the deterministic bottleneck model, and the stochastic bottleneck model in replicating actual airport traffic patterns. They show that the standard model cannot produce these rapid fluctuations.
on other vehicles. Travel demand is a decreasing function of the full price of travel, which consists of the travel time cost plus any toll for using the highway. The untolled equilibrium at the intersection of travel demand and the average travel-cost functions is socially inefficient because it does not account for the external cost of the trips. The optimal toll is the value of the vertical difference between the marginal and average travel times evaluated at the intersection of the travel demand and marginal social cost curves. It leads to the socially efficient traffic volume by internalizing the externality and equating marginal social costs and benefits of trips.

William Vickrey’s (1969) bottleneck model provides a dynamic model of congestion with travelers adjusting their time of travel to minimize their travel duration and schedule delay costs. Vickrey’s model uses a deterministic queue that develops at a highway bottleneck and prevents travelers from all arriving at their destinations at their most preferred times. The queue length depends on the entire traffic pattern starting from the most recent time it was empty and affects future travel delay until the queue is empty again. In the absence of tolling, equilibrium traffic and queuing patterns adjust endogenously over time to equate total costs of queuing and early or late arrival times of identical travelers. With deterministic queuing, optimal tolls adjust continuously throughout the peak period to mimic the queuing costs of the no-toll equilibrium. The congestion toll completely replaces the queue and converts all queuing costs into revenues. The model has several advantages over the standard model including: dynamic congestion, endogenous peaking of traffic and delay, modeling of traveler’s choice of when to travel, and accounting for schedule delays associated with travel time decisions. The model applies to peak-period pricing, and it has not previously been applied to airport slot constraints or other forms of quantity constraints.

The standard model is still the preferred framework among economists for modeling congestion. The economics literature largely ignored the bottleneck model until Arnott, de Palma, and Linsey (1990) reintroduced it by formalizing it, extending it to include a single-step toll, and modeling the optimal bottleneck capacity. Arnott et al. (1993) determines the optimal uniform, single-step, and continuous tolls with elastic demand. They argue that applying the standard model to subintervals of the peak period is conceptually unsound, but that the standard model represents a “semi-reduced form” of the entire peak period. They show that the efficiency gains from peak spreading are substantially greater than gains estimated with the standard model. They verify that the self financing result that Mohring and Harwitz (1962) and Strotz (1965)
proved for the standard model also applies to the bottleneck model, independently of the pricing regime.\textsuperscript{3}

Daniel (1991) develops a deterministic bottleneck model with congestion that is nonlinear in traffic rates and applies it to airport congestion. Daniel (1995) develops a bottleneck model with stochastic queuing and includes atomistic and non-atomistic traffic of Nash- and Stackelberg-dominant airlines. He implements the stochastic bottleneck model empirically using tower log data from Minneapolis-St. Paul airport and performs specification tests that indicate Northwest Airline’s scheduling of its flights is largely inconsistent with internalization of self-imposed delays. Daniel (2001) adds elastic demand, heterogeneous aircraft costs, and fringe aircraft with uniformly-distributed preferred operating times to the stochastic bottleneck model. Daniel (2009b) develops a model of multiple-step tolling for a deterministic bottleneck with dominant and fringe traffic. It uses a similar framework to determine the efficient tolling rules when airlines choose their operating times optimally in response to tolls, queuing delay, and schedule delay.

A moderate sized literature examines the economics of airport slot constraints. Much of the literature focuses on alternative methods of distributing slots, see e.g., Borenstein (1988), Starkie (1994), Mehdiratta and Kiefer (2003), Condorelli (2007), Verhoef (2008), and Brueckner (2009). A number of articles address the technical details of designing slot auctions; see Grether, Isaac, and Plott, C. (1989), Rassenti, Smith, and Bulfin (1991), de Vries and Vohra (2000), Jehiel and Moldovanu (2003). The political economy of implementing slot constraints has also received attention from Rose (2003), and Condorelli (2007), and Winston and de Rus (2008). Of these, only Verhoef (2008) and Brueckner (2009) explicitly model airport congestion and the optimal allocation of airport slots between dominant and fringe airlines. Verhoef demonstrates that in the presence of market power and uninternalized congestion, the first best pricing differentiates the tolls for dominant and non-dominant airline. Tradeable slots dominate undifferentiated pricing and approach the optimum when market power is unimportant, but slots may be less efficient than non-intervention when market power is significant. Brueckner (2008) includes congestion pricing, slot sales, and tradeable slots, but assumes away market power. He

\textsuperscript{3} Ralph Braid independently extends the bottleneck model to cover elastic demand. Arnott et al., (1989) and Yuval Cohen (1987) extend the bottleneck model to heterogeneous travelers.
shows that direct slot sales by the airport fail to achieve the optimum allocation because they have a uniform price. Dominant airlines have too little traffic volume and fringe airlines have too much. When slot quantity is fixed under a slot auction or slot trading, congestion is also fixed and dominant carriers can bid up the price of slots so that the equilibrium is efficient. Verhoef and Brueckner use a standard congestion technology for which delay is a function of current period traffic volume, so the models do not capture changes in the timing of operations within the period.

This paper contributes to the literature by developing a bottleneck model in which airports choose the timing and quantity of their slot-constraints, subject to dominant and fringe traffic that adjusts its time of operation within the slot windows to minimize its schedule- and queuing-delay costs. The bottleneck model raises and answers important questions that cannot be formulated within the standard model. These include the effects of intertemporal schedule adjustments on the optimality of slot constraints, how distribution of slots between airlines over time affect efficiency, how the duration of slot windows affects efficiency of equilibria, how slot constraints affect the ability of dominant carriers to internalize self-imposed delays, and the effect of slot constraints on strategic behavior by dominant airlines to preempt their most preferred periods.

The issues about how to specify the quantity and timing of slot permits are more basic than issues about how to design mechanisms to distribute them among airlines. The literature has largely overlooked these basic issues, but unless policymakers address them, the resulting slot-constraint system may have little effect on congestion at most airports, no matter how elaborately they design the slot auctions.

4. The model

First consider an airport with limited runway capacity facing demand for landings and takeoffs at rates that exceed the airport’s service rate. Air traffic control alternates landing and takeoffs in such a manner that landing and takeoff queues operate independently of each other.4 A generic model applies separately to both landings and takeoffs so the parameter names are the same, but the parameters values are different. The airport serves both a dominant airline that may

---

4This is approximately true of actual airports operating under balanced traffic (landing and takeoff) conditions. The author's observations of traffic counts indicate that somewhat higher rates of takeoff are possible when there are no landings, but that no additional landings are possible when there are no takeoffs.
coordinate its traffic to internalize delays that its aircraft impose on one another, and fringe airlines that operate their aircraft independently of one another (atomistically). The dominant airline operates a hub-and-spoke network and schedules \( d = d[p_d] \) aircraft per busy period, where \( p_d \) is the full price of a dominant aircraft operation, which is determined below. The dominant airline chooses a number, \( x \), of its \( d \) aircraft to schedule atomistically, while it internalizes the delays that its remaining \( d-x \) aircraft impose on one another. The fringe airlines operate \( f = f[p_f] \) aircraft per busy period, all operating independently of one another. Airlines would like to schedule their aircraft during a particular peak period to operate at the same time \( t^* \), which is the beginning or ending of the dominant airline’s passengers interchange period at its hub airport and is also a particularly desirable time for the fringe airline passengers to travel. \(^5\) Runway capacity limits the landing and takeoff rates to \( s \) aircraft per minute. The rate at which aircraft join the queue at time \( t \) is \( r[t] \). Deterministic queues develop at the runway bottlenecks from time \( t_{ab} \), when the queue is empty, according to the equation:

\[
q[t] = \int_{t_{ab}}^{t} r[u]du - s(t - t_{ab}).
\]

Because of limited capacity, aircraft must operate before or after their preferred times at costs \( \beta \) or \( \gamma \) per minute. Time spent in the queue costs \( \alpha \) per minute. Private landing or takeoff costs for atomistic aircraft are:

\[
C[t] = \alpha \frac{q[t]}{s} + \beta \max \left[ 0, t^* - t - \frac{q[t]}{s} \right] + \gamma \max \left[ 0, t + \frac{q[t]}{s} - t^* \right].
\]

In equilibrium, atomistic traffic rates adjust to maintain constant costs \( C[t] = C_i^* \) among groups of identical aircraft across all times in which they operate. The subscript on \( C_i^* \) denotes the slot window and indicates an important difference between slot-constrained equilibria and tolled equilibria. Costs are different across slot windows because slot constraints prevent traffic from moving between windows, whereas tolling leaves aircraft free to shift between toll periods resulting in the same cost throughout the entire peak period. \(^6\) Within each slot window, however,

\(^5\) The model can have preferred operating times that are distributed over time, but this complicates the solution without fundamentally changing the traffic patterns or policy implications.

\(^6\) It follows that the optimal slot constraint equilibrium is different from and better than the optimal tolled solution, but for reasonable parameters the difference is quite small.
costs are constant over all times in while atomistic aircraft actually operate. Solving for $r[t]$ separately when $t$ is early, $t \leq t^* - q[t]/s$, or late, $t > t^* - q[t]/s$, by substituting (1) into (2), and differentiating with respect to $t$ while imposing the constant cost condition gives the aggregate traffic rates for periods in which atomistic aircraft operate.\(^7\) The interior solutions for atomistic traffic rates are:

$$r[t] = \begin{cases} \frac{a}{a-\beta} s, & \text{for } t \leq t^* - \frac{q[t]}{s}, \text{ and} \\ \frac{a}{a+\gamma} s, & \text{for } t > t^* - \frac{q[t]}{s}. \end{cases}$$

Equations (1)-(3) are the basic framework of the bottleneck model due to Vickrey (1969) and Arnott et al. (1990, 1993).

4.1 Extending the model: optimal multiple-window slot constraints with atomistic traffic

Proposition 1: Optimal slot-constraint systems redistribute the flow of traffic during peak periods to reduce the accumulation of aircraft in the queues. Airports cannot directly control the traffic rate within slot windows, but they can effectively turn traffic flows on and off. They configure the slot windows and quantities of permits to periodically turn the flow on and allow the queue to fill, then turn the flow off and allow the queues to empty. By reducing the accumulation of aircraft in the queue, they reduce the total amount of queuing delay. Slot constraint systems increase in effectiveness as the number of slot windows increases and the duration of the slot windows narrows.

To derive the optimal slot-constraint system for a given number of periods, the airport authority chooses the quantity of operations to permit during each slot window and the beginning and ending times, $t_1$ and $t_2$, of the central window. The airport may also choose an offset $t_\varepsilon$ for $t_1$ and $t_2$, that shifts the entire busy period earlier or later relative to the most preferred operating time, $t^*$, to trade off total (social) queuing delays against schedule delays. The central period begins $t_1 + t_\varepsilon$ minutes before and ends $t_2 - t_\varepsilon$ minutes after the most preferred operating time, $t^*$. The durations of all the other periods follow from the number of slot permits issued per period and the values of $t_1$, $t_2$ and $t_\varepsilon$. For the purpose of solving the optimization problem, it is easier to

\(^7\) In the extended model, there are also periods with corner solutions when no traffic joins the queue because the cost is above the equilibrium level, or when a finite amount of traffic simultaneously joins the queue because the cost is lower than any other available operating time.
define the chosen periods in terms of when aircraft complete service. The actual slot windows that correspond to these periods are defined in terms of when aircraft arrive for service.

The simplest way to set up the problem of optimizing multiple-window slot constraints is by reference to the graph in Figure 1, which shows time on the horizontal axis and the number of aircraft operations on the vertical axis. The most preferred operating time is normalized to zero, so that negative times represent early operations and positive times represent late operations. The vertical axis shows the number of early and late aircraft using the same sign convention. The cumulative-service-completion function is a straight line through the origin with slope $s$, the service rate. The airport chooses slot quantities and time periods, but the airlines choose the traffic rates according to Equations (3). Consequently, the central window works like an unconstrained bottleneck equilibrium, except that the airport chooses when it begins and ends and the fraction of aircraft that operate during it. During the early and late parts of the central peak, the traffic rates must be $\alpha s/(\alpha - \beta)$ and $\alpha s/(\alpha + \gamma)$, so that the constant cost condition holds within each individual window. The connected line segments above and to the left of the service completion function indicate the cumulative number of aircraft arriving at the queue. The horizontal distance between the cumulative-arrival and service-completion functions is the aircraft queuing time. The area of the triangle defined by the cumulative arrival and service functions is the total queuing time. The length of the period and quantity of slot permits determines the size of the triangle. Simple geometry determines the queuing time of early aircraft operating during the central window to be $\beta(t_1 + t_2) t_1 s/(2\alpha)$ and that of those operating late to be $\gamma(t_2 - t_{1o}) t_2 s/(2\alpha)$. The triangle below the service completion function and to the left of $t^*$ has area equal to total early time, and the triangle below and to right has area equal to the total late time. Using simple geometry, the total early time delay during the central window is $(t_1 + t_{1o}) t_1 s/2$ and the total late time delay is $(t_2 - t_{1o}) t_2 s/2$.

Moving to the periods to the left of the central window, the airport chooses the quantity of slot permits to issue for each slot window. With deterministic queuing, it is always desirable to use every service interval during the busy period, so choosing slot quantities is equivalent to choosing the durations of the periods. The rate of arrival for service must equal $\alpha s/(\alpha - \beta)$ or zero during the early windows to satisfy the equilibrium condition of constant (minimum) cost within each window. Suppose that the airport chooses a slot quantity of $q_i$ for a window of duration $q_i s/2$ minutes. If the window is defined in terms of permissible service completion times, then as soon
as the period starts the traffic rate is $\alpha s/(\alpha-\beta)$ until all the permitted aircraft have joined the queue. At this point the queuing time is exactly equal to the remaining time in the window—operating later would result in missing the slot window. No more operations are desirable or possible for the remainder of the period, by which time the queue returns to its state at the beginning of the period, i.e., empty.

For late aircraft, a group of operations mass at the beginning of the slot window because the queue is empty and there is no benefit to waiting. These operations increase the queue to the length necessary to raise the expected cost of aircraft in the group to the slot window’s equilibrium level. The actual queuing delay of aircraft in this group ranges from zero to twice the equilibrium level. The cost of operating immediately after this group of aircraft is greater than the equilibrium level. If the value of queuing time exceeds that of late time, then queuing costs diminish fast enough to reach the slot window’s equilibrium level, at which point aircraft resume operating at the rate $\alpha s/(\alpha+\gamma)$. The fraction of aircraft in the initial mass is $\min\{1, 2 \gamma/(\alpha+\gamma)\}$ and the fraction operating later is $\max\{0, \gamma(\alpha-\gamma)/(\alpha(\alpha+\gamma))\}$. Again, the total queuing time for these periods is the area of the triangles formed by the cumulative arrival and service functions. The areas of these triangles is clearly $(\min\{1, 2 \gamma/(\alpha+\gamma)\} q_i^2/(2s)) + (\max\{0, \gamma(\alpha-\gamma)/(\alpha(\alpha+\gamma))\} q_i^2/(2s))$. The schedule delay of aircraft in period $i$ is the areas of the triangles and rectangles below the cumulative service and to the left of $t^*$. This value is also the number of aircraft in each period $q_i$ multiplied by the average waiting time from $t^*, q_i/(2s)+\sum_{j=1}^{i} q_j/s$.

Now it is possible to explain the role of $t_\varepsilon$ more completely. The choice of $t_\varepsilon$ shifts the entire peak relative to $t^*$. This shift enables the airport authority to start the peak later relative to $t^*$ so that the fraction of early to late aircraft varies, which is not possible by varying only $t_1$ and $t_2$, because of the constraints, $t_1 s + t_2 s = q_0$ and $\beta(t_1 + t_\varepsilon) = \gamma(t_2 - t_\varepsilon)$. To accomplish this, add or subtract $t_\varepsilon$ to $t_1$ and $t_2$ in the horizontal (time) dimension, but not the vertical (aircraft) dimension. This adjustment, however, is only possible when the value of late time exceeds the value of queuing time ($\gamma > \alpha$) so that all the late aircraft arrive en masse at the beginning of their slot window. Since no aircraft operate later in the window, the costs of operating in those times do not have to equal the cost at the beginning of the period. This frees the airport authority to shift the entire peak. Empirical estimates show that airport queuing time is more costly than late time ($\alpha > \gamma$), so $t_\varepsilon$ is constrained to be zero in this application. For the sake of generality, the optimization problem includes $t_\varepsilon$, but its first order condition is replaced by the constraint $t_\varepsilon = 0$. 

12
The airport authority chooses \( t_0, t_1, t_2, q_y, \ldots, q_z \) to minimize the total cost of aircraft operations in a multiple slot-constraint system with a given number of early slot windows, \( y \), and late slot windows, \( z \), plus the central slot window, \( 0 \), and subject to the constraint that all aircraft are served, \( f+x=m \). Its objective function is:

\[
\begin{align*}
\text{Minimize} & \quad t_1, t_2, t_e, q_{-y}, \ldots, q_z \left( \frac{\alpha \beta (t_1+t_e) t_1 s}{2 \alpha} + \frac{\alpha \beta (t_1-t_e) t_2 s}{2 \alpha} + \frac{\beta (t_1+t_e) t_1 s}{2} + \frac{y(t_2-t_e) t_2 s}{2} + \\
& \quad \sum_{i=1}^{y} \left( \frac{a \beta (q_{-i})^2}{2 s} + \beta \left( t_1 + t_e + \sum_{j=1}^{i-1} \frac{q_{-j} + q_{-i}}{s} \right) (q_{-i}) \right) + \\
& \quad \sum_{i=1}^{z} \left( \frac{\min \{2, 2 r q_i/(\alpha + y)\}^2}{2 s} + \alpha \frac{\max \{2 r (\alpha - y) q_i/(\alpha + y)\}^2}{2 s} \right) + \gamma \left( t_2 - t_e + \sum_{j=1}^{i-1} \frac{q_j + q_i}{s} \right) (q_i) \right) - \\
& \quad \lambda_1 \left( \sum_{i=-y}^{z} q_i - m \right) - \lambda_2 (t_1 s + t_2 s - q_0) - \lambda_3 (\beta (t_1 + t_e) s - y (t_2 - t_e) s) \right).}
\end{align*}
\]

The general solution to (4) is found by solving for many specific numbers of early and late slot windows \( y \) and \( z \), guessing the pattern of solutions as \( y \) and \( z \) change, and verifying that this general solution satisfies the first and second order sufficient conditions of (4). Let \( r \in [-y, z] \) denote a particular slot window. Define the parameters:

\[
\begin{align*}
\xi_{yz} &= 4 \gamma (2 \alpha^2 - 2 \alpha y + y^2) (2 \beta (1 + y) + y (2 + y)) + \alpha (\alpha + y)^2 (\beta + 2 \gamma + (\beta + \gamma) y) z, \\
\zeta_{yz} &= (2 (4 \gamma (2 \alpha^2 - 2 \alpha y + y^2) + \alpha (\alpha + y)^2 z) (1 + y))/(2 (y + 1)), \text{ and} \\
\psi_{yz} &= 4 (\beta + \gamma) (2 \alpha^2 - 2 \alpha y + y^2) + 4 \beta (2 \alpha^2 - 2 \alpha y + y^2) y + \alpha (\alpha + y)^2 z.
\end{align*}
\]

The general solution for the slot-constraint system when \( \alpha > \gamma \) is:

\[
q_{y,z,r} = \Gamma_{y,z,r} m, \quad \text{where} \quad \Gamma_{y,z,r} = \begin{cases} 
\frac{\gamma}{(y+1)(\beta+\gamma)}, & \text{for } r < 0, \\
\frac{\gamma \psi_{yz}}{(y+1)(\beta+\gamma) \zeta_{yz}}, & \text{for } r = 0, \text{ and} \\
\frac{a \beta (\alpha + y)^2}{(\beta + \gamma) \xi_{yz}}, & \text{for } r > 0.
\end{cases}
\]

\[
t_{y,z,1} = \frac{\gamma^2 \psi_{yz}}{(y+1)(\beta+\gamma)^2 \zeta_{yz}} \frac{m}{s}; \quad t_{y,z,2} = \frac{\beta \gamma \psi_{yz}}{(y+1)(\beta+\gamma)^2 \zeta_{yz}} \frac{m}{s}; \quad \text{and} \quad t_{n, e} = 0.
\]

The minimized value of the total cost function is:

\[
tc = \Gamma_{y,z,tc} \frac{\beta \gamma m^2}{s}, \quad \text{where} \quad \Gamma_{y,z,tc} = \frac{\xi_{yz}}{2 (\beta + \gamma) \zeta_{yz}}.
\]
When \( y \) and \( z \) are zero, the model produces the same total cost function, \( \beta \gamma (\beta + \gamma) \cdot m^2/s \), as the untolled and uniform-tolled equilibria of Arnott et al. (1993). If the airport imposes no quantity constraint (or ineffective constraints), the quantity demanded equals the unconstrained equilibrium quantity. If the airport authority optimally limits the quantity for the single slot window, then the equilibrium is the same as the uniform-tolled equilibrium, except that the initial slot owners keep the value of the un-assessed toll. In the limit as \( y \) and \( z \) go to infinity, the model produces the same total cost function, \( \beta \gamma / (2(\beta + \gamma)) \cdot m^2/s \) as the fine toll of Arnott et al. The multiple-slot window model covers the entire range of precision in slot constraint systems in a general unified model. To optimize the total number of aircraft operations, the airport authority should choose the total number of slot permits such that the marginal social benefit (demand) equals the marginal social cost. For example, the atomistic equilibrium with linear demand, \( m = \eta - \pi p_{y,z}^e \), is fully characterized by Equations (5), (6), (7) and:

\[
\begin{align*}
(8) \quad p_{y,z}^e &= \Gamma_{y,z,tc} \frac{\delta m_{y,z}^s}{s}; \\
m_{y,z}^s &= \frac{s \eta}{s + \Gamma_{y,z,tc} \delta \pi}; \\
t_{ab} &= t^* - \frac{y}{\beta + \gamma} \frac{m_{y,z}^s}{s}; \quad \text{and} \quad t_{ae} = t^* + \frac{\beta}{\beta + \gamma} \frac{m_{y,z}^s}{s}.
\end{align*}
\]

The solution up to this point reveals the general structure of optimal slot-constraint systems without the additional complication of strategic interaction between the airport and dominant and fringe aircraft. Figure 2 illustrates the equilibrium when there are three early and two late slot windows, plus the central window surrounding the most preferred operating time. During the build up to the peak, each window serves the same fraction, \( \gamma / (\gamma + 1)(\gamma + \beta) \), of the total aircraft in the busy period. If the airport authority redefines the slot windows in terms of when the aircraft arrives for service, then each early window should begin when the queue reaches the maximum desired length. The arrival rate drops to zero as aircraft wait for both the queue and schedule delay to decrease. Just as the queue empties, the arrival rate jumps to \( \alpha s / (\alpha - \beta) \) and the queue builds until the end of the slot window when all aircraft with a slot permit for the period have joined the queue. The next slot window starts and the arrival rate drops to zero as the process repeats. The central slot window starts when the queue from the previous period reaches its maximum, but aircraft wait until the queue empties at \( t_1 + t_c \) to begin arriving at rate \( \alpha s / (\alpha - \beta) \). The shift from early to late service completion times causes the traffic rate to drop to \( \alpha s / (\alpha + \gamma) \) so that the queue gradually empties until time \( t_2 - t_c \). Late slot windows start when the queue is empty and a fraction, \( \beta \gamma / ((\eta \gamma + \alpha)(\gamma + \beta)) \), of all the aircraft permitted to
operate during the period join the queue at the same instant. Air traffic control randomizes their service order to ensure that there is no incentive for them to try to beat the rush. After these aircraft simultaneously join the queue, the cost of additional operations exceeds the equilibrium cost from Equation (2). No more traffic occurs until the cost drops to the equilibrium level. At this point, traffic resumes at the late arrival rate, \( \alpha s/(\alpha + \gamma) \), as the queue slowly empties and the slot window ends. In Figure 2, the early slot windows begin at the left end of the flat regions of the cumulative arrival functions and the late slot windows begin just before the vertical regions. The number of slot permits in each window is the vertical difference in cumulative arrivals over the slot window.

Effective slot constraints work by issuing groups of slot permits with common slot windows so that modest queues develop, followed by periods with no new traffic that allow the queue to empty. The point of this optimal slot pattern is to reduce the accumulation of aircraft in the queue during the early periods. Slot constraints defer permission to join the queue until later service intervals when the traffic rate would otherwise be too low. The effectiveness of this peak spreading improves as the number of slot windows increases. Table 1 shows how the efficiency of the slot-constraint system varies with the number of early and late slot windows. The table uses cost parameters, \( \alpha, \beta, \) and \( \gamma \), that are typical of those Daniel and Harback (2008) estimate for major hub airports in the US, but the overall efficiency results in the table are not particularly sensitive to variations in the parameters. The cost parameters do affect the relative advantage of early versus late slot windows. Three slot windows with the optimal quantities of permits recover half of the efficiency loss from congestion, but eight slot windows are necessary to recover eighty percent. Such slot windows would be about ten minutes long for a typical peak.

4.2 Non-optimizing Slot Constraints

The above model includes the case with the optimal quantity of slots for a single slot window covering an entire peak period. This slot quantity is less than the unconstrained traffic level and results in a smaller peak. To achieve this, the timing of the slot window has to assure that the queue is empty when the constrained peak begins. The slot window must be at least as wide as the unconstrained peak to prevent aircraft from participating in the constrained peak by operating just before the slot window begins.\(^8\) It follows that the number of permits for the slot

\(^8\) The exact length of the window depends on demand in the period before the unconstrained peak, which is not
window is less than the airport’s capacity, and should be set according to Equation (6). While it is possible to have an effective system with single-window constraints, actual and proposed slot systems are generally characterized by wide time windows with slot quantities equal to the maximal number of operations that the airport is consistently able to perform over the duration of the slot window. This type of slot constraint motivates the following proposition:

**Proposition 2:** Suppose the airport authority chooses a slot window that covers an entire unconstrained atomistic traffic peak and sets the number of slots equal to its service capacity. With fully atomistic traffic, such slot constraints are non-binding and the result is equivalent to the unconstrained equilibrium.

Slot constraints with time windows that cover entire peak periods cannot spread atomistic traffic within the peak periods, so they cannot improve the efficiency of aircraft scheduling. Moreover, the total traffic volume over the complete peak period is equal to the service capacity, so this type of quantity constraint does not reduce the amount traffic that may participate in the peak. It follows that such slot constraints are completely ineffective for atomistic traffic. Marketability of these slot permits cannot improve their efficiency because they confer identical operating rights, so market prices cannot differentiate between operating times.

Proposition 2 illustrates the fundamental conceptual problems with the standard congestion model. When congestion is a function of current period traffic rate, it cannot model how the timing of operations within the peak period affects delays. The standard model attempts to account for fluctuation in traffic rates by slicing periods into shorter intervals. This approach is problematic because the standard congestion technology ignores the effect of traffic in adjacent periods on current congestion. In the standard model, wide time windows imperfectly capture the effects of proximate traffic on current delays but ignore congestion caused by traffic fluctuations within slot windows. Narrow time windows partially control for these traffic fluctuations but ignore the spillover congestion from adjacent periods. Real world slot constraints typically have one hour windows, while traffic rates may repeatedly fluctuate above and below the capacity rate within the hour. Such slot-constraint systems do little to reduce congestion, but they may serve as barriers to entry by potential competitors (Dresner, et al., 2002). Studies of traffic and congestion

---

9 There may be the right number of aircraft already in the queue when the window starts so that it empties exactly when the constrained peak begins, and no capacity is unused.
patterns at US and Canadian airports by Daniel and Harback (2009) and Daniel (2009a) demonstrate that significant congestion persists at slot-constrained airports like Toronto’s Pearson International and occurred at previously slot-constrained airports like Washington National, LaGuardia, and J. F. Kennedy.

4.3 Unconstrained equilibrium with dominant and fringe aircraft

Proposition 3: In the unconstrained dominant-fringe equilibrium, the dominant airline fully internalizes when fringe demand is perfectly inelastic; partially internalizes when demand is imperfectly elastic; and approaches fully atomistic behavior as the fringe approaches perfectly elastic demand. The dominant airline internalizes delays by scheduling some aircraft to operate at the service rate just before and after the atomistic peak.

Now it is necessary to distinguish between the numbers of dominant and fringe aircraft, aircraft. Let number of dominant and fringe aircraft in the atomistic peak be $x$ and $f$, and the number of internalizing dominant aircraft be $d-x$. Fringe aircraft do not care about the identity of other aircraft operating at about the same time they do, so Equation (3) describes the aggregate traffic rates even when dominant aircraft operate during the atomistic peak. Dominant aircraft would not exceed the fringe rates during the peak period because doing so is more expensive than adopting the atomistic rates. Therefore, the best (scheduling) responses of the atomistic fringe as functions of the dominant airline’s arrival rates are:

$$ r_f[t; r_d[t]] = \begin{cases} 
0, & \text{for } C[t] > C^* \\
\max \left[ 0, \frac{\alpha x}{\alpha - \beta} - r_d[t] \right], & \text{for early periods,} \\
\max \left[ 0, \frac{\alpha x}{\alpha + \gamma} - r_d[t] \right], & \text{for late periods, and} \\
\infty, (\text{finite simultaneous arrivals}), & \text{for } C[t] < C^*.
\end{cases} $$

The fringe best response function in Equation (9) shows that the dominant airline cannot reduce its aircraft costs by rescheduling them within the atomistic peak. To internalize self-imposed delays, the dominant airline must shift some aircraft off peak. The least costly rate at which to schedule these aircraft is the service rate $s$, because they cannot obtain service more rapidly than rate $s$ and there is no further advantage to schedule aircraft at rates below $s$. The operating times of the first and last dominant aircraft, $t_{db}$ and $t_{de}$, must be such that their costs are identical, otherwise the airline could reduce costs by moving aircraft from the higher- to the lower-cost period. It follows that the dominant airline’s traffic pattern for aircraft not scheduled
during the atomistic peak is:

\( r_d[t] = s, \) for \( t \in [t_{db}, t_{ab}] \) and for \( t \in [t_{ae}, t_{de}] \), where

\[
t_{db} = t^* - \frac{\gamma}{\beta + \gamma} \frac{d + f}{s} \quad \text{and} \quad t_{de} = t^* + \frac{\beta}{\beta + \gamma} \frac{d + f}{s}.
\]

There are \( \gamma / (\beta + \gamma) \cdot (d-x) \) early aircraft that experience early time of \( (d+x+2f) \gamma / ((\beta + \gamma)2s) \) and \( \beta / (\beta + \gamma) \cdot (d-x) \) late aircraft that experience late time of \( (d+x+2f) \beta / ((\beta + \gamma)2s) \). Multiplying the numbers of aircraft by their time values and average delay times gives the total cost of the internalizing dominant aircraft, \( \delta(d-x)(d+x+2f)/(2s) \).

Recall from Equation (8) that the minimum total, marginal, and average cost functions are of the form \( \Gamma \delta (d^2)/s, 2\Gamma \delta m/s, \) and \( \Gamma \delta m/s \) where \( \Gamma \) is the constant specified in Equation (7). In this subsection, there are no slot constraints so there is one unconstrained period, \( \Gamma \) is equal to one, and after substituting \( (f+x) \) for \( m \), the full price is \( \delta(f+x)/s \). Fringe demand determines the number of fringe aircraft based on the full price that the fringe aircraft experience. Substituting the full price of fringe operations in the fringe demand function and solving for equilibrium quantity yields:

\[
(11) \quad p^e_f[f,x[d]] = \frac{\delta(f+x[d])}{s}; \quad \text{and} \quad f^e_s[x[d]] = \eta - \pi \frac{\delta(f^e_s[x[d]]+x[d])}{s} \Rightarrow f^e_s[x[d]] = \frac{\eta s - \pi \delta x[d]}{s + \pi \delta}.
\]

Adding the total cost of the dominant aircraft in the atomistic peak and substituting the fringe demand gives the dominant airline’s objective function. The dominant airline’s problem is to choose the number of aircraft to schedule during the atomistic peak to solve:

\[
(12) \quad \text{Minimize} \quad x \left\{ \frac{\delta (d-x)((d-x)+2(f^e+\pi))}{2s} + \frac{\delta (f^e+\pi)x}{s} \right\}, \quad \text{s.t.} \quad 0 \leq x \leq d, \quad \text{and} \quad \hat{f}^e = \frac{\eta s - \delta \pi x}{s + \delta \pi}.
\]

It is easy to see that the solution to this problem is to choose the number of aircraft to schedule atomistically as follows:

\[
(13) \quad \hat{x}^e[d] = \frac{\delta \pi d}{s + \delta \pi}, \quad \text{for} \quad \frac{\delta \pi}{s + \delta \pi} \geq 0.
\]

Let \( \phi = \delta \pi / (s + \delta \pi) \) denote the fraction of dominant aircraft scheduled during the atomistic peak. Substituting \( \phi d \) for \( x[d] \) in the dominant airline’s average cost yields the airport supply function,
which determines the full price for dominant operations in unconstrained equilibria. Let the supply and demand functions for dominant airline operations be:

\[ p^e_d[f, d] = \frac{s}{s} + \frac{d}{s}(1 + p^2) \quad \text{and} \quad d^e[p^e_d] = \mu - \rho p^e_d[f, d]. \]

The number of aircraft in the unconstrained equilibrium simultaneously satisfies the supply and demand functions as given in Equations (11) and (14). Let \( \hat{f}^e \) and \( \hat{d}^e \) be the number of aircraft in the untolled equilibrium. Let \( \hat{p}^e_f \) and \( \hat{p}^e_d \) be the equilibrium full prices of fringe and dominant aircraft. The reduced-form solutions for unconstrained equilibrium prices and quantities are:

\[ \hat{f}^e = \frac{n}{\phi} (\delta \rho s(\phi^2 + 1) + 2 s^2) - \frac{n}{\phi} (2 \delta \pi s) \quad \text{and} \quad \hat{p}^e_f = \frac{s}{s} (\hat{f}^e + \phi \hat{d}^e); \]
\[ \hat{d}^e = \frac{\mu}{\phi} (2 \pi s + 2 s^2) - \frac{\mu}{\phi} (2 \delta \rho s) \quad \text{and} \quad \hat{p}^e_d = \frac{s}{2s} (2\hat{f}^e + (1 + \phi^2)\hat{d}^e), \]

where \( Y = 2 s^2 + \delta^2 \pi \rho (\phi - 1)^2 + \delta s(\rho(\phi^2 + 1) + 2 \pi). \)

Note that \( \phi \) is the fraction of aircraft that the dominant airline schedules atomistically as claimed in Proposition 3. This fraction equals zero when the slope of the fringe demand curve is zero and approaches one as the slope goes to minus infinity. Equation (10) describes the behavior of the \( 1 - \phi \) internalizing dominant aircraft which is also as claimed in Proposition 3.

4.4 Optimal discrete slot constraint systems with dominant and fringe aircraft

Proposition 4: If the airport authority can fully determine when the dominant and fringe aircraft operate, then it will separate their operations to enable the dominant airline to fully internalize. The airport either reserves all the peak-period slot permits for the fringe and the dominant airline fully internalizes by scheduling all aircraft off-peak, or it allocates all central peak slots to the dominant airline and the fringe schedules its aircraft on either side of them. If the airport is unable to enforce this separation, then the dominant airline will schedule some of its aircraft atomistically during the peak periods as in the unconstrained equilibrium.

Both the airport authority and the dominant airline are potentially strategic players now. Different assumptions about their strategies or objectives lead to different equilibria, but three sets of assumptions seem particularly defensible. The simplest solution is for the airport authority to allocate a fixed number of peak-period slots to the fringe as specified in Equation (6). The dominant airline perceives fringe demand as inelastic and schedules all its aircraft off peak at
exactly the service rate. Alternatively, the airport authority could allocate all central peak-period slots to the dominant airline to schedule at the service rate while the fringe schedules aircraft atomistically during the surrounding periods. Under the assumptions of homogeneous time values and airlines choosing the traffic rates, these are socially optimal outcomes for the given distribution of slot windows. The airport authority must know the dominant and fringe demand functions to determine the optimal number of slots to allocate. Either the dominant or fringe airlines would find these outcomes objectionable because they are confined to the least desirable slots. Both would prefer to have the peak-period slot permits. The third possibility assumes that the airport authority has less control over airline schedules and will make peak slots available to the fringe if the dominant airline it does not use them. When fringe demand for peak slot permits is not constant, the dominant airline attempts to preempt these slots by scheduling some aircraft atomistically. This solution is similar to the unconstrained equilibrium, except that the slot constraints reduce fringe demand elasticity and increase dominant internalization.

Equation (6) determines the timing of the slot windows during the atomistic peak and the optimal fraction of traffic to assign to each slot window as functions of the number of peak operations \((f+x)\). Equation (7) specifies the social cost of these atomistic aircraft operations. Adding the social cost of the internalizing dominant aircraft gives the social cost of all aircraft operating during the busy period:

\[
C_s = \left[ \delta (d-x) \left( \frac{(d-x)}{2s} + \frac{f+x}{s} \right) + \Gamma \frac{\delta (f+x)^2}{s} \right]
\]

Differentiating (16) with respect to \(f\), determines the marginal social cost of fringe operations:

\[
\frac{\partial C_s}{\partial f} = \frac{\delta (d-x)}{s} + \frac{2 \Gamma \delta (f+x)}{s}.
\]

The airport authority equates the marginal social cost of fringe operations with the marginal social benefit given by its inverse demand function:

---

10 The model assumes an isolated peak period with no earlier or later demand for operations. Slot permits for the off-peak aircraft may be necessary in real world applications to protect these periods from additional entry. It may also be necessary to limit traffic for a period before the off-peak aircraft to assure the queue is empty when the busy period begins.
\[
\eta - f = \frac{\delta (d - x)}{s} + \frac{2 \Gamma \delta (f + x)}{s} \Rightarrow \hat{f}^s[x, d] = \frac{\eta s - \delta \pi d - (2\Gamma - 1)\delta \pi x}{s + 2\Gamma \delta \pi}.
\]

Suppose that the authority has the upper hand and is able to control the division of dominant aircraft between atomistic and internalizing behavior. The airport authority chooses the number of dominant aircraft that behave atomistically, \(x\), to minimize social costs by setting the derivative of Equation (16) with respect to \(x\) equal to zero, substituting \(\hat{f}^s[x, d]\) from Equation (18) into the result, and solving for \(\hat{x}[d]\):

\[
\frac{\partial C_s}{\partial x} = \frac{(2\Gamma - 1)\delta (f + x[d])}{s} = 0 \Rightarrow \hat{x}[d] = \frac{\delta \pi d - \eta s}{s + \delta \pi}.
\]

Note that \(\hat{f}^s[x, d]\) and \(\hat{x}[d]\) cannot simultaneously both be positive, so non-negativity constraints require that \(\hat{x}[d]\) be zero whenever there is fringe traffic in the peak. The airport authority would not schedule dominant aircraft among the fringe aircraft when the alternative is to have them fully internalize their delays. Now the authority chooses the number of (internalizing) dominant operations by setting their marginal social cost equal to their marginal social benefit and substituting \(\hat{f}^s[x, d]\) from Equation (18) to solve for \(\hat{d}\). Finally, it substitutes \(\hat{d}\) back into (18) to obtain the reduced-form solutions for the total number of slot constraints to issue:

\[
\frac{\partial C_s}{\partial d} = \frac{\delta (f + d + (2\Gamma - 1)(f + x[d])x'[d])}{s} = \frac{\mu - d}{\rho} \Rightarrow
\]

\[
\hat{d} = \frac{s(\mu - \delta \eta \rho + 2\Gamma \delta \mu \pi \rho)}{s^2 + (2\Gamma - 1)\delta \rho \pi + \delta (\rho + 2\Gamma \pi)} \quad \text{and} \quad \hat{f} = \frac{s(\eta + \delta \eta \rho - \delta \mu \pi \rho)}{s^2 + (2\Gamma - 1)\delta \rho \pi + \delta (\rho + 2\Gamma \pi)}.
\]

To implement this slot constraint system, the airport authority makes available a total of \(\hat{f}\) slot permits to the fringe to operate during the atomistic peak. It divides them among the slot windows as specified in Equation (6). The airport authority could also issue the dominant airline \(\gamma \hat{d}/(\beta + \gamma)\) and \(\beta \hat{d}/(\beta + \gamma)\) slot permits to operate immediately before and after the atomistic peak. These permits are only necessary if additional demand for off-peak periods requires that the airport reserve them for the dominant airline.

Figure 3 illustrates the effect of slot constraints on the timing of dominant and fringe traffic. The thick dark line represents the unconstrained traffic pattern of the atomistic peak that

\footnote{The social benefits of dominant operations are unaffected by whether they internalize or not.}
is similar to the atomistic bottleneck equilibrium. Congestion externalities lead atomistic traffic to exceed the service rate during the early periods, because it ignores the effect of queuing delays it imposes on other traffic. The traffic rate falls below the service rate exactly when new operations complete service late. This allows the queue to empty at exactly the end of the peak. When a dominant airline internalizes self-imposed delays, it moves operations out of the atomistic peak to the adjacent regions on either side, as indicated by the light gray regions. This traffic operates at the service rate and creates no queuing. The solution in Equation (13) determines the fraction of dominant traffic remaining in the atomistic peak. The darker shaded regions illustrate the slot-constrained traffic rates. The early atomistic traffic rate still exceeds the service rate because it must still satisfy the bottleneck equilibrium condition in Equation (2). The airport cannot effectively set the traffic rates, but it can use the slot constraints to halt the traffic flow periodically to permit the queue to empty. Slot constraints spread the traffic more evenly over the peak—not by reducing traffic rates but by creating gaps and shifting traffic to low traffic periods. Slot constraints may also reduce the overall number of aircraft operations and affect the distribution of dominant aircraft between the atomistic peak and the dominant internalizing periods. In the equilibrium of Equation (20), the airport authority would shrink the atomistic peak by moving all dominant operations to the gray regions. In the next scenario, however, the dominant airline may leave some traffic in the atomistic peak to preempt fringe operations.

Now suppose that the dominant airline chooses the number of aircraft to schedule atomistically to minimize its private costs subject to the fringe demand under the slot constraint system. Equations (16)-(18) describe the social cost and the fringe behavior as before. The dominant objective function differs from the social cost in Equation (16) in that the \((f+x)^2\) in the last term is replaced by \((f+x)x\) because the dominant airline is only concerned with its own costs. The dominant airline accounts for the fringe’s best response to its schedule as in Equation (9), but the fringe now faces the full price given by Equation (7). The dominant airline’s problem and its optimal choice of \(x\) are:

\[
\begin{align*}
\text{Minimize} & \quad x \left\{ \delta (d - x) \left( \frac{(d-x)}{2s} + \frac{(f+x)}{s} \right) + \Gamma \frac{\delta (f+x)x}{s} \right\}, \\
\text{s.t.} & \quad \hat{f}^s[x, d] = \frac{\eta_s - \delta \pi d -(2 \Gamma -1) \delta \pi x}{s+2 \Gamma \delta \pi}, \\
\Rightarrow & \quad \hat{f}^s[d] = \frac{(3\Gamma -2)\pi \delta d -(\Gamma -1)x \eta}{(2 \Gamma -1)(s+2 \pi \delta)}, \\
\Rightarrow & \quad f^s[d] = \frac{s^2 \eta + s \delta (-d+\eta + \Gamma \eta) \pi -3d \Gamma \delta^2 \pi^2}{(s+2 \pi \delta)(s+2 \Gamma \delta \pi)}.
\end{align*}
\]
The full price and the number of dominant operations are:

\[
\begin{align*}
\text{atc}_d^2 &= \frac{\delta(s^2(-1+2\Gamma)(d+\eta)+\delta s\Gamma(d(-2+4\Gamma)+(-1+3\Gamma)\eta)\pi-\Gamma^2\delta^2\pi^2)}{s(-1+2\Gamma)(s+2\delta\pi)(s+2\Gamma\delta\pi)} \\
\tilde{d}_d^2 &= \frac{\Gamma^2(1-3\Gamma)\delta^2\eta\pi + (-1+2\Gamma)(2s^2(1+\Gamma)\delta\mu + s(s^2\mu + s\delta\eta + 4\Gamma^2\delta^2\mu\pi^2))}{-\Gamma\delta^2(s(4-8\Gamma)+2s\pi)^2 + (-1+2\Gamma)(s^2(s+\delta\rho)+2s\delta(s+\delta\Gamma+\delta\rho)\pi)}.
\end{align*}
\]

This equilibrium has mixed dominant and fringe traffic during the atomistic peak and spans the whole range of possibilities, from fully internalizing behavior to fully atomistic behavior (in the limit) by the dominant airline as it perceives the fringe demand responsiveness varying from perfectly inelastic to perfectly inelastic. In addition, as the number of slot windows varies from one to infinity, \( \Gamma \) ranges from one to one half. As Table 1 shows, the social cost of the operations varies from 200% to 100% of the socially efficient level. Efficiency appears to improve rapidly with the addition of slot windows. Three windows, one window on each side of central period, recover half the efficiency loss from congestion. Five early and three late windows added to the central window recovers eighty percent of the loss. Most real world slot constraint systems have only one window per hour, while most arrival or departure peaks last one-half to one and one-half have hours. Roughly speaking, recovering eighty percent of the efficiency loss would require five- to ten- minute slot windows. Even recovering only fifty percent of the loss would require ten- to thirty-minute slot windows. The questions that slot advocates must address are: “How will the system cope with the administrative costs of distributing (by whatever means) highly differentiated slot permits?” or alternatively; “Are the efficiency gains (if any) obtainable with hourly slot window worth the still substantial administrative costs of inadequately differentiated slot permits?” Clearly, the WSG system involves high administrative costs by requiring about one thousand people to attend biennial slot conferences to haggle over slot distributions for three days.

4.5 Continuous Slot Constraints:

As the number of slot windows in Equations (21) and (22) approaches infinity (or more realistically one per service interval) the atomistic aircraft fully internalize the social cost of their operations, eliminating the distinction between internalizing and atomistic behavior. The airport authority can treat the “internalizing” dominant aircraft the same as all the “atomistic” aircraft because it no longer matters when the dominant aircraft operate during the busy period. Different treatment of dominant and fringe aircraft is unnecessary. The easiest way to demonstrate this is
to use the explicitly continuous model. While slot constraints with this is high level of specificity are probably infeasible, the continuously varying toll schedule would also be the same for dominant and fringe aircraft and would probably be no more expensive to administer than the current weight-based pricing.

Proposition 5: The (first-best) optimal slot-constraint system has one slot permit and window per service interval. The optimal number of operations is the same as in the untolled equilibrium because the full price of operations is the same in either case. All the resource savings are due to rescheduling operations rather than reducing traffic levels. Whoever is initially endowed with the slot permit captures all of the deadweight loss from queuing that would have occurred in the absences of slot constraints.

The optimal continuous slot-constraint system is the limit of the discrete model as the number of early and late periods tends to infinity. This first-best social optimum eliminates all queuing and converts all deadweight loss into social surplus. The limit of $\Gamma_{y,z,tc}$ as $y$ and $z$ go to infinity is one-half, so the cost given by Equation (7) goes to $\delta m^2/(2s)$. Alternatively, the continuous slot-constrained equilibrium cost can be derived directly. It must have $m$ aircraft scheduled to operate at the service rate $s$ from time $t_{eb}=t^*-y/(\beta+\gamma)\ m/s$ until $t_{ae}=t^*+\beta/(\beta+\gamma)\ m/s$. The corresponding social cost consists of the total schedule delay costs. Clearly, the average early time $y/(2(\beta+\gamma))\ m/s$ times the value of early time $\beta$ multiplied by the number of early aircraft $\gamma(\beta+\gamma)m$ plus the average late time $\beta/(2(\beta+\gamma))\ m/s$ times the value of late time $\gamma$ multiplied by the number of early aircraft $\beta/(\beta+\gamma)m$ equals $(\beta+\gamma)/(2(\beta+\gamma))\ m/s \ \delta m = \delta m^2/(2s)$.

The marginal social cost for the optimal constrained equilibrium, $\delta m/s$, also equals the queuing cost of the aircraft with the longest queue in the unconstrained equilibrium, or the schedule delay cost of the first and last aircraft in the peak. The constraints begin at time $t^*-\beta/(\beta+\gamma)\ m/s$ and end at $t^*+\beta/(\beta+\gamma)\ m/s$, permit $s$ aircraft per minute. The traffic rate equals the service rate throughout the entire busy period and no queue develops. In Figure 2, the cumulative arrival function for the fine toll is coincident with the cumulative service completion function.

Slot constraints are often thought to be simpler to implement than congestion tolls, particularly fine tolls that vary continuously. In principle, a market for tradeable slots can solve the optimal tolling problem without needing the airport authority to set tolls. All the airport authority has to do is print slot certificates, choose any method of initial distribution, and let the market go to work. Proposition 5, however, says that to fully optimize the allocation of airport
capacity requires separate a slot window and permit for each individual service interval. The market cannot differentiate prices for slot permits unless they have different time windows. When slots are so highly differentiated, there are obvious practical problems in efficiently distributing slots, coordinating schedules, and verifying compliance. Congestion tolling, on the other hand, conveys the relevant information about marginal social costs, allows free entry, has relatively low administrative costs, and is error tolerant because queuing will clear the market if tolls are erroneous. The policy section, supra, considers these issues more thoroughly.

5. Policy discussion and conclusions

Proponents of slot-constraint systems argue that assessing different congestion prices for dominant and fringe aircraft is first-best optimal, but that price differentiation is politically difficult and imposes onerous information requirements. Auctions or markets for slot permits improve on undifferentiated congestion pricing, because dominant airlines out bid the fringe and capture more slots, without the airport authority needing to set prices. These arguments are based on the standard congestion model for which adjusting traffic volumes is the only means of internalizing delays. The deterministic bottleneck model, with its dynamic congestion and endogenous traffic rates, is more applicable to airport traffic because it allows airlines to internalize delays by rescheduling aircraft. In a dynamic model, slot permits must have different operating-time windows to support different prices. Airports cannot prevent traffic from peaking within slot windows. Slot-constraint systems face an inherent conflict between requiring many narrow slot windows to achieve reasonably efficient aircraft scheduling and few slot windows to simplify their distribution and administration costs. More slot windows means vastly more permutations of slot permits for the system to allocate.\(^\text{12}\) Narrow slot windows also make it difficult to verify that an aircraft has the proper slot permit when it randomly deviates from the permitted operating time.

The standard model ignores these dynamic issues by optimizing the number of aircraft operating within a few wide slot windows. This approach would still be reasonable if

\(^{12}\) The value of a slot permit depends on combining it with other slots. Within a single airport, slot portfolios must coordinate aircraft arrival and departure times with connecting aircraft and gate availability. With multiple airports, slot portfolios must coordinate arrival and departure times throughout the route network. Airlines must also coordinate slot combinations across time, by days, months, and seasons to provide coherent schedule frequency.
internalizing congestion were primarily a matter of reducing the total traffic volume rather than adjusting the timing of operations. The current weight-based pricing system, however, functions by creating queues that increase the full price of operating until demand equals supply at the current traffic volume. Optimizing the system involves replacing these queuing costs with congestion fees, slot prices, or quantity restrictions, not reducing the traffic level. In the deterministic bottleneck model, the optimal traffic level is the same in the unpriced or unconstrained equilibrium as the first-best equilibrium.

Calculating the marginal social costs of airport operations to determine the appropriate toll schedule is not really very difficult. In principle, the correct toll structure is simply the value of queuing-delay time under the untolled traffic pattern. The Federal Aviation Administration routinely calculates aircraft operating costs and passenger and crew time costs for use in cost-benefit analyses of airport improvement projects. It also collects information on airline fleet characteristics at all the major airports. These values and a correct understanding of the congestion technology are sufficient to determine the external delays. Daniel (1995) develops a stochastic bottleneck model and Daniel and Harback (2008) estimate time dependent airport delays from flight level data at major airports. They show that the stochastic bottleneck model matches the delay patterns very well, and they determine toll schedules for all major US airports.

Tolling solves or avoids a number of additional problems posed by slot constraints. The cost of administering time-dependant tolls should be comparable to the costs of administering the current weight-based pricing system. Slot constraints require a new expensive distribution system. Control towers already gather the information needed to asses time varying tolls, based on when aircraft join the arrival and departure queues. The tolls would depend on actual aircraft operating times that are readily observable. Slot constraints depend on scheduled operating times and require additional incentives to assure that airlines actually schedule aircraft (and operate them on average) at the permitted time. With tolling, airlines could schedule aircraft by checking the airport tolling structures and traffic patterns to determine the full price of operations, without having to obtain prior slot permits. If the toll levels are correct, there would be minimal queuing. If the toll levels are incorrect, queuing delay would equate supply and demand as it currently does. Persistent queuing would provide a signal to revise the toll schedule. New airlines would be free to enter an airport market as long as they valued an operation more than its social cost,
without obtaining prior permission. Slot-constrained airports show increasing concentration over time because of entry barriers. Airports would not necessarily need to differentiate prices between airlines or types of aircraft to optimize their traffic patterns. In the absence of price rationing, slots require administrative rationing with differential treatment of airlines. Tolling generates sufficient revenues to fund airport capacity, and revenues provide signals to indicate the need for additional capacity. Slots permits confer unpriced benefits on airlines because they are willing to pay more for operations than their full price. These benefits give rise to costly rent-seeking behavior. Finally, tolling allows more flexibility to adjust to changing conditions over time than slot constraints that depend on historical usage patterns or grant property rights lasting for a decade. If slot constraints were the only policy available to mitigate airport delays, then the policy’s difficulty in attaining the first-best optimum would be of little relevance. Congestion tolling, however, avoids these problems and requires only marginal changes in airport and airline practices. Daniel (2009b) develops a multi-step tolling model and discusses its implementation.

REFERENCES


Daniel (2009b) shows that a common step-toll schedule generally optimizes the scheduling of aircraft, while an additional uniform may be necessary to optimize the number of aircraft operating in a busy period.


<table>
<thead>
<tr>
<th>Early\Late</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>10</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>200</td>
<td>184</td>
<td>178</td>
<td>175</td>
<td>174</td>
<td>172</td>
<td>170</td>
<td>167</td>
<td>167</td>
</tr>
<tr>
<td>1</td>
<td>167</td>
<td>150</td>
<td>145</td>
<td>142</td>
<td>140</td>
<td>139</td>
<td>137</td>
<td>134</td>
<td>134</td>
</tr>
<tr>
<td>2</td>
<td>156</td>
<td>139</td>
<td>134</td>
<td>131</td>
<td>129</td>
<td>128</td>
<td>125</td>
<td>123</td>
<td>123</td>
</tr>
<tr>
<td>3</td>
<td>150</td>
<td>134</td>
<td>128</td>
<td>125</td>
<td>124</td>
<td>122</td>
<td>120</td>
<td>117</td>
<td>117</td>
</tr>
<tr>
<td>4</td>
<td>147</td>
<td>130</td>
<td>125</td>
<td>122</td>
<td>120</td>
<td>119</td>
<td>117</td>
<td>114</td>
<td>114</td>
</tr>
<tr>
<td>5</td>
<td>144</td>
<td>128</td>
<td>123</td>
<td>120</td>
<td>118</td>
<td>117</td>
<td>114</td>
<td>112</td>
<td>111</td>
</tr>
<tr>
<td>10</td>
<td>139</td>
<td>123</td>
<td>118</td>
<td>115</td>
<td>113</td>
<td>112</td>
<td>109</td>
<td>107</td>
<td>106</td>
</tr>
<tr>
<td>50</td>
<td>135</td>
<td>118</td>
<td>113</td>
<td>110</td>
<td>108</td>
<td>107</td>
<td>104</td>
<td>102</td>
<td>102</td>
</tr>
<tr>
<td>100</td>
<td>134</td>
<td>118</td>
<td>112</td>
<td>109</td>
<td>108</td>
<td>106</td>
<td>104</td>
<td>101</td>
<td>101</td>
</tr>
</tbody>
</table>

Note: Assumes 100 aircraft with $\alpha=50$, $\beta=7.5$, and $\gamma=15$. 
Figure 1—Geometric derivation of slot constraint equilibrium costs
Figure 2—Slot-constraint system with three early and two late slot windows
Figure 3—Comparison of traffic patterns with and without slot constraints