The Relationship Between Rationalized Schedules and Reduced Arrival Variability

John F. Shortle* and Lance Sherry*
Center for Air Transportation Systems Research
Dept. of Systems Engineering & Operations Research
George Mason University, Fairfax, VA 22030, U.S.A.

The objective of this paper is to determine the relationship between rationalized schedules and reduced arrival variability and their relative impact on reducing arrival delays. Many NextGen technologies and procedures aim to reduce arrival variability, while policies like congestion pricing and slot auctions seek to rationalize the schedule. Using a simple queueing model, we explore the relative impact of these reductions.

I. Introduction

According to classical queueing theory,1 queueing delays get worse as the variability in a system increases. In principle, it is possible to eliminate delays by eliminating variability. A perfectly synchronized (deterministic) system can be run with zero delays. For example, a highly synchronized subway system in which trains arrive at precisely spaced intervals might be a “near-deterministic” system. Typically, we think of variability as resulting from “randomness” or “noise” in the system. For example, in the National Airspace System (NAS), flights are often delayed because of uncertainties in weather, wind, maintenance, and other stochastic factors.

But variability can also arise from deterministic sources. For example, the arrival schedule (which is deterministic) may have peaks and valleys. This implies a variation in the scheduled arrival rate as a function of time throughout the day. Even though this variability arises from a deterministic source (the fixed schedule), it can still contribute to queueing delays. This occurs when the scheduled arrival rate exceeds the airport capacity over a short period of time. Queues develop during such periods even if all flights arrive at the destination airspace exactly on time. In the language of queueing theory, this type of variability reflects the non-stationary nature of the system.

The objective of this paper is to determine the relationship between these two types of variability. We refer to the first type of variability – the inherent noise and randomness in the system – as system variability. We refer to the second type of variability – the degree of unevenness in the underlying arrival schedule – as schedule variability. The objective is to determine the relative effects of reducing each type of variability and to identify the corresponding impact on queueing delays.

Many NextGen technologies and procedures aim to reduce system variability – for example, by using four-dimensional trajectories to reduce the variability in along-track flight time or by using improved systems to predict weather. The potential reductions in delays have been investigated by numerous researchers (for example,2–5). Policies like congestion pricing and slot auctions seek to reduce schedule variability by reducing the peaks in the schedule. The delay impacts of depeaking and similar policies have also been investigated in the literature (for example,6–11).

The contribution of this paper is to explore the relationship between these two types of variability. The work is exploratory in nature and the queueing models proposed in this paper are very simple in order to understand the key relationships.

* Associate Professor, Dept. of Systems Engineering & Operations Research, 4400 University Dr., MSN 4A6, Fairfax, VA 22030.
II. A Simple Delay Model

This section describes a simple queueing model to evaluate the relative effects of system variability and schedule variability. Before describing this model, we first define terms more precisely. Schedule variability is defined as the sample variance of the number of scheduled arrivals in disjoint $m$-minute intervals, where $m$ is a positive constant (e.g., $m = 15$ corresponds to 15-minute interval bins). To calculate the schedule variability, count the number of scheduled arrivals in one 15-minute interval (say, $X_1$), the number of scheduled arrivals in the next 15-minute interval (say, $X_2$), and so forth. The schedule variability is the sample variance of the sequence $X_1, X_2, \ldots, X_n$. A completely uniform schedule – that is, a schedule where the scheduled arrivals are spaced at precise intervals throughout the day – has a schedule variance equal to 0. A heavily banked schedule with large peaks and valleys has a larger schedule variance.

System variability is defined as the inherent uncertainty in the system that prevents an airplane from arriving to its destination on time. We define “arrival” time to be the time that an airplane enters the terminal airspace of its destination airport, rather than the landing “on” time or the gate “in” time. This definition is more consistent with queueing theory in which the arrival time is typically defined as the time that an arrival can first enter the queue (not the time that it actually starts service). In this context, the “arrival” time is roughly the time when the airplane enters the terminal airspace. In summary, schedule variability describes the variation in the schedule itself, and system variability describes an ability to meet the schedule.

The queueing model given in this paper is admittedly very simple and omits many components that would otherwise make it more realistic. Nevertheless, the objective is to capture the essential physics of the problem (queueing delays as a function of system variability and schedule variability) as efficiently as possible. Although the model is quantitative, the conclusions that can be drawn from the model are more qualitative in nature. The contribution is not to provide precise numerical estimates, but rather to give a mechanism to explain and understand the relationship between these types of variability and how they impact delays.

The model in this paper is a single-airport model. Thus, it does not provide system-wide estimates of delay. The model includes three basic elements: (1) A model for the arrival schedule, (2) a model for how flights deviate from the arrival schedule, and (3) a model for determining queueing delays based on arrival times. The model (described below) is completely specified by five parameters:

1. $\mu$: Airport arrival capacity (arrivals per hour),
2. $\lambda$: Average arrival rate (arrivals per hour),
3. $A$: Amplitude of schedule peaks (arrivals per hour),
4. $T$: Time period between arrival banks (hours),
5. $\sigma$: Standard deviation of system noise (minutes). This refers to the stochastic variability in the arrival time to the terminal airspace.

We describe these components in detail.

A. Schedule Model

Figure 1 shows the scheduled arrivals for Hartsfield-Jackson Atlanta International Airport (ATL) for a sample day in the summer of 2010. (A midweek day was chosen, but otherwise the choice of a particular day was somewhat arbitrary. The data source was the Aviation System Performance Metrics (ASPM) database. ASPM does not include canceled and diverted flights, so the scheduled counts in the figure may slightly underestimate the actual schedule.)

The main observation from the figure is that the schedule is somewhat oscillatory in nature. This reflects the bank operations of Delta and Air Tran. The arrivals and departures are synchronized to enable transfer connectivity between a large number of arriving and departing flights. The figure shows a sample “fit” of a sine function to the actual schedule. (The parameters of this particular function were not obtained in a rigorous manner, but simply by trial and error.) Of course, the sine function does not provide a very good fit for the whole day, but it does provide a reasonable match between noon and 8pm. This motivates the use of an oscillatory function, like sine or cosine, as a very simple model to capture the effects of schedule banks.
Figure 1. Arrival schedule at ATL, Thur. Jun. 24, 2010.

The sine wave in Figure 1 is parameterized by three values (see Figure 2) – amplitude \( A \), period \( T \), and average value \( \lambda \) – and can be written as follows:

\[
\lambda(t) = \lambda + A \sin(2\pi t/T), \quad \text{or} \quad \lambda(t) = \lambda + A \sin(\omega t).
\]

The parameters \( \lambda, A, \omega, \text{ and } T \) are constants, where the angular frequency \( \omega \) and the period \( T \) are related via \( \omega = 2\pi/T \). The function \( \lambda(t) \) is interpreted as the instantaneous scheduled arrival rate, meaning that the number of scheduled arrivals in a short time interval of length \( \Delta t \) is approximately \( \lambda(t) \Delta t \). It is assumed that \( A \leq \lambda \); otherwise, \( \lambda(t) < 0 \) for some values of \( t \).

In Figure 1, the fitted sine function has the following parameters: \( \lambda = 110, \ T = 1.36, \text{ and } A = 60 \). This means, to a rough approximation, the average arrival rate is 110 per hour, there are 1.36 hours between banks, and the peak arrival rate exceeds the average arrival rate by 60 per hour. Note that the \( y \)-axis in Figure 1 is in units of “per quarter hour” so the plotted function is \( \lambda(t)/4 \) rather than \( \lambda(t) \). As a reference point, the average utilization at ATL, assuming an average scheduled arrival rate of \( \lambda = 110 \) per hour, is somewhere above \( \rho \geq .9 \). (The maximum capacity at ATL, based on declared airport arrival rates obtained from ASPM data, is about 215.\textsuperscript{12} The optimal rate from the Airport Benchmark Report 2004 is 243, including planned improvements through 2013.\textsuperscript{13} Assuming an even split between arrival and departure capacity, \( 110/(243/2) \approx .9 \).)

Of course, not all airports have oscillatory schedules. For example, LaGuardia Airport (LGA) does not have a noticeable periodic structure in its arrival schedule. This is because LGA is slot controlled and does
not have a single dominant carrier. We can still apply the oscillatory model of (1) to this type of airport in a very rough way. The parameter $A$ (which must fall in the range $[0, \lambda]$) can be used to characterize the trade-off between two basic extremes: On the one extreme is a perfectly “flat” schedule (a slot constrained airport with evenly spaced scheduled arrivals) which is represented by $A = 0$. On the other extreme is a “pure bank” schedule (an airport with a single carrier running banked operations) which is represented by setting $A = \lambda$. The latter case maximizes the peaks and valleys relative to the average arrival rate. In particular, $\lambda(t) = 0$ in the valleys. While no real airport falls at either extreme, LGA is closer to the “flat” schedule ($A = 0$) and ATL is closer to the “pure bank” schedule ($A = \lambda$). More specifically, based on the fitted sine function in Figure 1, $A/\lambda \approx .55$ for ATL.

To generate simulated scheduled arrival times from the rate function $\lambda(t)$, define $\Lambda(t)$ as the cumulative number of scheduled arrivals by time $t$:

$$\Lambda(t) \equiv \int_0^t \lambda(x)dx = \int_0^t [\lambda + A \sin(\omega x)] dx = \lambda t + \frac{A}{\omega} [1 - \cos(\omega t)].$$  \hspace{1cm} (2)

$\Lambda(t)$ is a non-decreasing function of $t$ provided $A \leq \lambda$. The $n$th arrival is assumed to occur at the first instant $t$ such that $\Lambda(t) = n$. This is essentially a discretization of the continuous function $\Lambda(t)$.

B. Arrival Model

In this paper, “arrival” time (either the scheduled arrival time or the actual arrival time) refers to the time that a flight enters the terminal airspace near the airport. This corresponds to a time that a flight can first enter the landing queue. The actual arrival time is assumed to equal the scheduled arrival time plus a term representing random noise.

Actual arrival time = Scheduled arrival time + $N(0, \sigma^2)$,

where $N(0, \sigma^2)$ is a normal random variable with a mean of 0 and a standard deviation of $\sigma$.

A slightly more realistic model might be to use a gamma distribution instead of a normal distribution. The gamma distribution is non-symmetric: It allows for very late arrivals but excludes very early arrivals.

C. Queueing Model

The airport is assumed to operate as a single-server first-come-first-served queue with deterministic service times. In the case of a single runway, the service time $D$ can be thought of as the minimum time separation between successive arrivals. A reasonable value might be $D = 90$ seconds, corresponding to a maximum capacity of $\mu = 40$ arrivals per hour. Although the model assumes a single server, an airport with two arrival runways might be approximated by doubling the capacity $\mu$. In general, the capacity of the airport is specified by the parameter $\mu$ (arrivals per hour) and the deterministic service time is set internally as $D = 1/\mu$.

Again, this is a simple model in which many complicating factors are ignored. In addition to simplifying the airport as a single-server queue, the model also assumes that it is possible to space aircraft exactly $D$ seconds apart. In reality, there is stochastic variability on this separation. The observed standard deviation of this separation is around 15-20 seconds under a fully loaded arrival queue.\textsuperscript{14–17} In general, most assumptions in this paper can be easily generalized to more complicated scenarios and analyzed within a similar framework.

The constant $\lambda$ represents the average arrival rate. It determines, in part, the average utilization of the system. If the airport capacity (maximum number of arrivals per time) is $\mu$, then the average utilization is $\rho = \lambda/\mu$ or $\rho = \lambda D$. In the long run, the utilization must be less than 1. Otherwise, the queue is unstable and delays grow without bound over time. This does not preclude the possibility that the instantaneous arrival rate exceeds the maximum capacity. For example, in Figure 2, $\lambda(t)$ exceeds $\mu$ for brief periods of time. However, because the overall arrival rate $\lambda$ is less than $\mu$, the system is stable in the long run.

III. Methodology

Because of the non-stationary nature of the model, analytical results using classical queueing theory are not available. (Certain limiting results can be obtained analytically. For example, as $\sigma \to \infty$, the arrivals become “randomly” distributed in time. The model becomes an $M/D/1$ queue, for which analytical results
exist. Green and Kolesar have analyzed a similar $M(t)/M/c$ queue in which the non-stationary arrival rate is sinusoidal as in (1). In their model, since the arrival process is a non-homogeneous Poisson process and service times are exponential, the queue can be evaluated by numerically integrating the Chapman-Kolmogorov differential equations. In our model, neither the arrival process nor the service times are exponential, so we simply apply discrete-event simulation to analyze the queueing system. The system is simulated using the following steps:

1. Generate scheduled arrival times according to (2). Specifically, the arrival time of the $n$th flight is set equal to the value of $t$ such that $\Lambda(t) = n$.

2. Generate actual arrival times by adding an i.i.d. normal random variable to each scheduled arrival time. The actual arrival times are then sorted into ascending order.

3. Evaluate the average delay (in queue) using a first-come-first-served single-server queue. This can be evaluated quickly in a standard way using Lindley’s recursive equation.

We conduct a set of experiments by varying several system parameters. One of the system parameters to vary is the ratio of the schedule amplitude ($A$) to the average arrival rate ($\lambda$). This ratio is varied ($A/\lambda$) from 0 to 1 in increments of 0.1. $A/\lambda = 0$ corresponds to a completely flat (evenly spaced) schedule. $A/\lambda = 1.0$ corresponds in some sense to a “maximum bank size.” The ratio cannot exceed 1.0, because otherwise $\lambda(t)$ would be negative for some values of $t$; see (1). Also, the noise parameter is varied from $\sigma = 0$ (corresponding to a deterministic system) to $\sigma = 10$ minutes. For each combination of parameters, the system is simulated for 120 hours of simulated time. The time to simulate all 121 combinations ($11 \times 11$) is fairly short – less than one minute. The discrete-event simulation is written in C++.

**IV. Results**

Figure 3 shows sample results for a system where the average utilization is $\rho = 0.8$. The minimum separation is $D = 90$ seconds corresponding to a maximum capacity of $\mu = 40$ per hour. Thus, $\rho = 0.8$ implies that the average arrival rate is $\lambda = 32$ per hour. The bank period is set at $T = 1.36$ hours, corresponding to the sample ATL schedule (Figure 1).

![Figure 3](image.png)

**Figure 3.** Average delay per flight as a function of system variability ($\sigma$) and schedule variability ($A/\lambda$); $\rho = 0.8$.

The two graphs in Figure 3 show the same results, but the left graph gives a two-dimensional view and the right graph gives a three-dimensional view. The left graph shows average delay (per flight, in minutes) as a function of the schedule variability. Average delay can be converted to total daily delay (at one airport) by multiplying by the number of flights. For example, the total number of daily scheduled arrivals for ATL in Figure 1 is 1,397 (this is from ASPM, so excludes canceled and diverted flights). An average delay of, say, 3 minutes per flight corresponds to about 4,200 total minutes of total daily delay at the airport.

The $x$-axis of the left graph corresponds to schedule variability: $A/\lambda = 0$ corresponds to an evenly spaced schedule. $A/\lambda = 1$ corresponds to a maximally banked scheduled, and $A/\lambda = 60/110 \approx .55$ is the value obtained from the fitted sine function based on the ATL schedule in Figure 1. The graph shows different...
curves for different values of the system-variability parameter $\sigma$, where $\sigma = 0$ corresponds to a deterministic system. Note that $\sigma$ does not correspond to the standard deviation of aircraft spacing on final approach, which is on the order of 15-20 seconds.\textsuperscript{14-17} Rather, it represents the standard deviation of the arrival time to the terminal airspace of the destination airport, so it includes many stochastic factors like gate push-back delay, weather delays, mechanical delays, flight-time variability, and so forth.

One way to interpret the graph is as follows: Suppose that a certain level of system variability is estimated (say, $\sigma = 4$ minutes). This value might be regarded as an external fixed parameter, since it is difficult to control. Given this level of system variability, the degree of schedule variability that should be allowed in the system can then be determined. For example, if $\sigma = 4$ minutes, then the highest value of $A/\lambda$ that keeps average flight delays under 2 minutes is approximately $A/\lambda = 0.4$.

The right graph of Figure 3 shows the same information, but in three dimensions. In an ideal system, the schedule variability and system variability are both zero ($A/\lambda = 0$ and $\sigma = 0$). This is the left-most point in the three-dimensional figure. Such a deterministic system has zero delays. Of course, this is not realistic. An actual airport might be somewhere in the middle of the figure (e.g., $A/\lambda \approx .55$ based on the ATL schedule). While the figure shows a notional baseline point, we emphasize that the point should not be taken as a quantitative representation of any real airport due to the many simplifications in modeling.

The key observation from the two graphs is the following: The schedule variability has a more significant effect on delay than the system variability. Regardless of where the baseline point is located, it generally appears to be more advantageous to reduce the schedule variability (that is, to move along the $A/\lambda$ axis) than to reduce the system variability (that is, to move along the $\sigma$ axis).

Figure 4 shows analogous results for a system with lower average utilization, $\rho = 0.5$. In this figure, the system variability appears to have a more significant effect. That is, it can be more advantageous to reduce the system variability compared with the schedule variability; however, this conclusion depends on the reference location of the baseline point. For example, when $\sigma = 10$ (minutes), there is little benefit in a reduction to $\sigma = 6$ (this is seen more clearly in the left graph). However, a reduction from $\sigma = 6$ to $\sigma = 0$ yields a significant reduction in delays.

![Figure 4](image)

**Figure 4.** Average delay per flight as a function of system variability ($\sigma$) and schedule variability ($A/\lambda$); $\rho = 0.5$.

The qualitative difference between Figures 3 and 4 can be explained intuitively as follows. When $\rho$ is high, the effects of high peaks in the schedule dominate. They concentrate a large number of arrivals over a short period of time. When $\rho$ is high, this causes $\lambda(t)$ to be significantly higher than the capacity $\mu$ during the peaks, thus increasing delays. In some case, it may even be theoretically possible to decrease delays by increasing system randomness. That is, when the schedule peaks are severe, the noise in the system can actually help by spreading out the peaks. This effect is actually seen in Figure 3. In contrast, when $\rho$ is low, and in particular when $\rho \leq 0.5$, then the scheduled arrival rate $\lambda(t)$ never exceeds the capacity $\mu$, even when $A/\lambda = 1$. Thus, when the system operates deterministically ($\sigma = 0$), there are no delays regardless of the value of $A/\lambda$ (assuming $0 \leq A/\lambda \leq 1$). That is, the schedule itself does not create problems when $\rho$ is low. In this case, queueing only results because of uncertainty in flights to arrive at their destination on time.

However, we must be somewhat careful in making conclusions here. The figures do not show the amount of effort or cost in achieving a given reduction in either variable. The two horizontal axes correspond to distinctly different variables and the relative slopes of the two arrows in the graph can be changed simply by...
re-scaling either axis. For example, we can make either variable look more “important” simply by drawing the other variable over a smaller range of values. But what we can conclude is the following: For larger values of $\rho$, the relative importance of schedule variability increases. Further, the potential benefits in reducing system variability are limited unless there is also a simultaneous reduction in schedule variability.

Finally, we look at the impact of increasing airport capacity $\mu$. Procedures to allow more operations on closely spaced runways can improve capacity. Also, reducing the variability in spacing on final approach can indirectly improve capacity by allowing a reduced safe separation between aircraft. Figure 5 shows the impact of increasing $\mu$ (or equivalently, decreasing the schedule utilization $\rho = \lambda/\mu$, holding the scheduled arrival rate $\lambda$ fixed), based on our simple model. The two graphs show the pairwise combined effects of reducing $\rho$, system variability, and schedule variability. The main observation is that reducing $\rho$ (or increasing $\mu$) and reducing $A/\lambda$ seem to have larger relative impacts than reducing $\sigma$. This is particularly true when $\rho$ is high. Then a significant benefit is seen by either increasing $\mu$ (decreasing $\rho$) or decreasing $A/\lambda$. A related work analyzes the effects of bank sizes and queue utilization on delays.11

Figure 5. Effects of utilization, system variability, and schedule variability.

A. Arrival Profiles

In this section, we consider a more realistic arrival process based on an actual schedule (e.g., Figure 1) rather than a sinusoid function. To incorporate a schedule into the model, define $\hat{\lambda}(t)$ as the actual number of scheduled arrivals during a given 15-minute interval, normalized to an hourly rate. For example, based on Figure 1, $\hat{\lambda}(t)$ would be the same as the schedule shown in the figure, except that it would be multiplied by 4 to convert to an hourly rate and it would be regarded as a continuous function of time (that changes every quarter hour).

Let $\lambda$ be the average scheduled arrival rate over a given time interval – that is, the total number of scheduled arrivals in an interval divided by the number of hours in that interval. We restrict the time interval of interest to the busy period of the day, say, from 8am to 8pm, so that $\lambda$ is not excessively biased by the quiet overnight period. The model assumes that time runs continuously from the end of one day to the start of the next, so by “skipping” the night hours the model effectively eliminates the downtime that the system might otherwise use in reality to catch up.

Similar to (1), define

$$\lambda(t) = \lambda + \frac{A}{c}(\hat{\lambda}(t) - \lambda), \quad \text{where} \quad c = \lambda - \min\{\hat{\lambda}(t)\}. $$

The proportionality constant $c$ is chosen so that $A$ is used analogously as in (1). The parameter $A$ controls the “flatness” of the hypothetical schedule. Figure 6 illustrates. When $A = 0$, the schedule is completely even; $\lambda(t) = \lambda$ (approximately 22 arrivals per quarter hour). As $A$ increases, the variability of the schedule increases. It increases in a way that maintains the shape of the original schedule relative to the average value. When $A = c$, the original schedule is achieved; $\lambda(t) = \hat{\lambda}(t)$. The maximum value of $A$ is $\lambda$, corresponding to the point where increasing the variability further would result in negative values for $\lambda(t)$. This method of
adjusting the schedule does not change the width of the banks or shift operations from peak-period banks to off-period banks as might be done in reality.

The cumulative arrival function $\Lambda(t)$ is defined as the integral of $\lambda(t)$; see (2). Since $\hat{\lambda}(t)$ is a piecewise constant function, $\Lambda(t)$ is a continuous piecewise linear function. In summary, the main assumptions of the model extension are the following: (a) the number of scheduled arrivals in each interval is proportional to the difference between the original scheduled arrival rate and the average arrival rate, (b) the proportionality constant $A$ corresponds to the variability of the schedule ($A = 0$ corresponds to a flat schedule, $A = c$ corresponds to the original schedule, and $A = \lambda$ corresponds to a “maximum-variability” schedule), and (c) scheduled arrivals are assumed to be evenly spaced over each 15-minute interval.

Figure 7 shows delay results using the extended model. The service rate $D$ is set so that $\rho = 0.8$. Thus, Figure 7 is analogous to Figure 3. From a qualitative perspective, the figures are very similar. Although the ranges of the vertical axes are slightly different, the qualitative shapes of the graphs are similar. In particular, it still appears to be more advantageous to reduce the schedule variability compared with the system variability.

**V. Conclusions**

This paper explored the relationship between rationalized schedules and reduced arrival variability using a simple queueing model. Key insights from the model are:
1. When utilization is high, a rationalized schedule (or reduced schedule variability) is a significant component of reducing delays.

2. When utilization is low, low arrival (or system) variability is a significant component of reducing delays.

3. Either way, there is a non-linear relationship between schedule variability and system variability and also airport capacity. Changing any one of these elements effects the relative impact of the other two. Thus, a delay-benefits analysis of any new technology or procedure that impacts any one of these elements should also consider the potential impacts of changes to the other elements.

The model presented in this paper is relatively simple. There are many avenues that can be explored to extend the results. These include obtaining estimates for baseline parameters based on data, using a better queueing model for the airport (e.g., multiple runways and more complex runway configurations), and building correlation into the noise structure (e.g., during bad weather, each flight may be affected in a similar manner, so the noise adjustments are not independent).

References