Simulating Airport Delays and Implications for Demand Management

Vikrant Vaze

December 7, 2009

Contents

1 Operational Irregularities and Delays 3

2 Motivation for a Delay Simulator 4

3 The M|G|1 Simulator with a 'Schedule' 5
   3.1 Arrival process ............................................. 5
   3.2 Service time distribution .................................... 7
   3.3 Number of servers ........................................... 9
   3.4 Implementation details ....................................... 9

4 Dynamic Aspects of Queuing 11
   4.1 Delays when capacity exceeds demand ...................... 11
   4.2 Impact of GDP ................................................... 12
   4.3 Delay peak often lags behind the demand peak ............. 13
   4.4 Marginal delay cost as a function of time-of-day .......... 13

5 Implications for Demand Management 15
   5.1 Quantity based demand management methods ............... 17
   5.2 Price based demand management methods ................... 19
      5.2.1 Algorithm 1: Alternate .................................... 21
      5.2.2 Algorithm 2: Alternate with moving averages ........ 22

6 Summary of Findings and Key Takeaways 23
List of Figures

1. Causes of NAS delays .......................... 4
2. Effect of service time variation on expected total delay ........ 9
3. Effect of service time variation on variability of total delay .... 10
4. Tradeoff between run-time and delay variance ............... 11
5. Delays when capacity exceeds demand .................. 12
6. Impact of capacity reduction ........................ 13
7. Delay peak lags behind the demand peak ................. 14
8. Marginal delays as a function of time-of-day ............. 16
9. Total delay reduction due to decrease in allocated capacity .... 19
10. *Alternate* algorithm does not converge ............... 22
11. *Alternate with moving averages* algorithm converges very fast ... 23

List of Tables

1. Undesirable effects of pure Poisson assumption .............. 8
2. Delay reduction under quantity based demand management .... 20
1 Operational Irregularities and Delays

The airport infrastructure is a critical element of profitability and success of the airline industry. For decades, airlines and air traffic controllers and managers have focused on several operations research techniques for minimizing costs, maximizing revenues and improving safety of various components of aviation system. Traditional approaches of improving profitability have concentrated on increasing the planned profit. As air travel increased over the years, the costs due to delays have skyrocketed. Airlines and airport managers now realize that optimizing the planned costs is not enough, operational deviations from the planned schedule do play a significant role in actual realized costs. During the 4-5 years after the 9/11 attacks, the delays costs rose steeply. While the current economic recession has undermined their impact for the last year or so, many studies indicate that large delays will make a return as the recession subsides and passenger demand increases. Therefore, it is important to understand the main drivers of airport delays and take proactive measures to lessen the impact of delays before strong economic conditions resume.

As compared to the original plan of operation, airline schedules often deviate significantly on the day of operation due to delays and disruptions. These operational irregularities necessitate reactive measures called the airlines recovery process. There is a significant additional cost associated with such irregularities. This cost is in addition to the planned operating cost. According to Air Transport Association’s estimates [ATA08], the total additional aircraft operating cost due to delays in United States was $8.1 billion in the year 2007. Sherry and Donohue [PDR08] estimated the total cost of delays to the passengers to be $8.5 billion for the same year. It is interesting to note that the total operating profits of all the US domestic carriers were about $4.4 billion in that year [BTS09]. This gives an idea of the extent of delay problem in the United States. According to the website of Bureau of Transportation Statistics [BTS09], delays to 50% of the delayed flights were attributable to the National Aviation System (NAS). Further, figure 1 illustrates the distribution of NAS delays by cause for year 2007 [BTS09], which shows that 84.5% of the NAS delays were attributed to either volume or non-extreme weather conditions. In other words, most of the NAS delays were attributable to either scheduling more operations than maximum capacity or to the reduction in realized capacity because of inclement weather conditions. Both these effects have a common theme in that the delays are caused by scheduling more operations than the realized capacity. So demand-capacity mismatch can be identified as the single most important driver of aviation delays.
2 Motivation for a Delay Simulator

Most commercial aviation flights in United States follow planned schedules. However, the actual arrival times of the flights are typically distributed around this scheduled arrival time. Upon arrival into an airport’s terminal airspace, aircraft join a queue and then they are allowed to land one-by-one depending on the stipulated separation requirements for safety. There are several factors that determine the actual time required for the landing of an aircraft including runway configuration, weather conditions, type of aircraft, type of previous aircraft, air traffic controllers’ skill level, pilot skill level, terrain and other airport specific factors etc. So the aircraft landing durations can also be described by a probabilistic distribution. Therefore, arrivals at an airport can be modeled as a queuing system. In some cases, the aircraft are subject to ground holds. So instead of queuing in the airspace near an airport, they queue up at various locations around the country. Even this is an example of a spatially distributed queuing system.

Queuing theory recognizes different queuing models based on different assumptions about the distributions of arrivals and service times and the number of servers. Closed form solutions are available for some of these systems under the steady-state assumptions. But in reality steady state conditions are rarely, if ever, observed at the congested airports, because the number of scheduled arrivals varies significantly across the day and the queuing system rarely gets a chance to
get stabilized in a steady state. Dynamic behavior of queues can be significantly different from their steady state description. Sometimes cumulative diagrams are used to study dynamic queues. But cumulative diagrams fail to take into account the probabilistic flavor of these problems. So the two alternative ways of solving such systems are through simulation or through numerical solution of system of equations.

In this project, we will adopt the simulation based approach. We intend to use simple models that simulate airport arrivals with a reasonable degree of accuracy. The main aim is to understand and appreciate the various aspects of the relationship of delays with demand and capacity. Another objective is to test the effects of some demand management techniques and to study their computational properties. We develop a simple simulator for flight arrivals at an airport. Section 3 describes the simulator in details. We have learnt various concepts about the dynamic nature of queues in this class. Section 4 describes the results when the simulator is used to test some of these concepts. In section 5, we will use this tool to investigate the impacts of two different demand management techniques. Finally, we will summarize the findings in section 6.

3 The M|G|1 Simulator with a 'Schedule'

In the queuing system terminology, M|G|1 implies that the arrivals at the queuing system are described by a Poisson/Memoryless (M) process, the service times are given by general (G) distribution and the number of servers is 1. We will look at each of these assumptions one by one and discuss them in terms of implementation.

3.1 Arrival process

In a Poisson process, the inter-arrival times have negative exponential distribution. This means that there is a finite nonzero probability associated with the inter-arrival times falling in any finite length interval over the non-negative numbers. Further, the probability that the inter-arrival times exceed any nonnegative number is also finite, regardless of how large the number is. Though these assumptions sound unrealistic, the probabilities of such weird events are extremely low. This makes the exponential distribution a good candidate for simulation of flight arrivals in the terminal airspace. Another interesting feature of Poisson process is that given that a certain number of arrivals happen in a certain time period, the unordered arrival time of each of these arrivals follows a Uniform Distribution. So one way of simulating the arrival process is to simulate the number of arrivals in a time period using Poisson distribution and then simulate the actual arrival times as a uniform distribution. Let us call this way of simulation as the Pure Poisson
Simulation. While this is an accurate way of simulating the Poisson process of arrivals, it may not be the most realistic for flight arrivals that follow a schedule. Given that a certain number of flights demand service during an hour, it is almost impossible that the actual number differs substantially from this number. Even if the actual number is different, this would mean that the number of flights arriving in the subsequent intervals is likely to be affected by how different this number is. Another practical constraint is that the total number of flights demanding landing service at an airport during an entire day is practically constant. In order to take these practical issues into consideration, we have modified the arrival simulations process as follows. We will divide the entire day into finite number of intervals of 60 minutes each. Next we will assume that the number of arrivals in that interval is the same as the scheduled number. Finally, we will simulate the actual arrival times based on a uniform distribution. From here onwards we will denote this way of simulation as Poisson with Schedule.

In order to substantiate the issues with the pure Poisson assumption, we compared the results under the two assumptions. Delays were simulated at the LaGuardia Airport (LGA) in New York based on the actual demand profile observed at LGA on 31st January 2008 under VFR conditions. Assuming Pure Poisson process, the delays were simulated. The true demand profile was used as the demand rate to simulate the actual number of arrivals in each time period from a Poisson distribution. Additionally, another simulation run was carried out under the Poisson with Schedule assumption. We also observed that the run-time under the Poisson with Schedule assumption was around 50 seconds while that under the Pure Poisson assumption was 117 seconds, which makes sense since the Pure Poisson simulation requires more computations.

Total delay in the Pure Poisson case was found to be 3533 minutes while that in the Poisson with Schedule case was 1936 minutes. Obviously, since the Pure Poisson case is more random, it makes sense that the delay should be higher. However, the delay was found to be almost 82% higher than the Poisson with Schedule case. This would be a cause for concern. Clearly the two assumptions would lead to extremely different values of delays, which means that at least one of the two assumptions is fairly inaccurate. So further analysis of intermediate computations was carried out. Table 1 summarizes the details of this analysis.

The Pure Poisson assumption leads to too much randomness. As seen in table 1, the minimum simulated demand for each of the 24 time periods was 0 while the maximum demand exceeded the airport capacity significantly in all the time periods from 7 am to 11 pm. Also as the table suggests, the variance of demand was very high. The maximum hourly demand was as high as 72 in some cases whereas the capacity was merely 40.5. Further, the maximum daily demand was observed to be 721 arrivals, while the minimum was 0 arrivals. We can compare these numbers
to the scheduled daily demand of 594 arrivals. In reality, the number of arrivals per day can be assumed to be constant. In some cases, flights may get cancelled, but having additional 127 arrivals compared to the daily scheduled demand is very much impossible. The scheduled number of operations at the airport during many of the hours already is close to the capacity. Therefore, the delays are very sensitive to demand variations at this level of demand. We are able to understand why the total delay under the Pure Poisson assumption is so drastically more than the Poisson with Schedule case. However the kind of demand changes per hour as summarized in 1 just don’t make sense realistically. Therefore, from here onwards, we will abandon the Pure Poisson assumption and work under the Poisson with Schedule case for the rest of this paper.

3.2 Service time distribution

The M|G|1 queuing system can accommodate any general probability distribution of service times. We will use the capacity values published by FAA under the VFR and IFR conditions to estimate the average service times. In order to introduce some stochasticity in service times, we will assume that the service times can vary within 5% of the average value and the distribution is assumed to be uniform.

The introduction of stochasticity into the service times should increase not only the variance of delay but also the expected value of delays. The rationale for increasing variance of delay is quite straightforward. More the variability in service time, more is going to be the variability in delays. However, the reason why the expected delay value should increase due to increase in variance of service times is less obvious. The key fact is that the delays are nonlinear functions of utilization ratios and the second derivative of expected delay with respect to utilization ratio is positive. In other words, with increase in utilization ratios, the rate of change of expected delay increases. An increase in service time would lead to a linear increase in utilization ratio under constant demand and a decrease in service time would lead to a linear decrease in utilization ratio. But the increase in expected delay due to increase in utilization ratio is more than the decrease in expected delay due to decrease in utilization ratio by the same magnitude. So under symmetric perturbation of service times, the expected service time stays constant but the asymmetric variability in expected delays means that there is going to be a net increase in expected delay. This is something that can be tested using this simulator.

Figure 2 shows the change in expected delay and figure 3 shows the change in standard deviation of delay. On x-axis of both charts is the percentage of average service times within which the service times are allowed to vary uniformly. For example, a point corresponding to x value of 20% in either of these charts means
Table 1: Undesirable effects of pure Poisson assumption

<table>
<thead>
<tr>
<th>Hour</th>
<th>Scheduled Demand</th>
<th>Capacity</th>
<th>Demand Variance</th>
<th>Maximum Demand</th>
<th>Minimum Demand</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>40.5</td>
<td>2.00</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>40.5</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>40.5</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>40.5</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>40.5</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>40.5</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>40.5</td>
<td>3.98</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>36</td>
<td>40.5</td>
<td>36.11</td>
<td>63</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>34</td>
<td>40.5</td>
<td>33.99</td>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>38</td>
<td>40.5</td>
<td>38.11</td>
<td>69</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>35</td>
<td>40.5</td>
<td>35.27</td>
<td>64</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>35</td>
<td>40.5</td>
<td>35.03</td>
<td>63</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>39</td>
<td>40.5</td>
<td>38.92</td>
<td>66</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>36</td>
<td>40.5</td>
<td>36.28</td>
<td>68</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>34</td>
<td>40.5</td>
<td>34.21</td>
<td>62</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>36</td>
<td>40.5</td>
<td>35.93</td>
<td>67</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>38</td>
<td>40.5</td>
<td>37.85</td>
<td>68</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>38</td>
<td>40.5</td>
<td>37.87</td>
<td>66</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>34</td>
<td>40.5</td>
<td>34.15</td>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>37</td>
<td>40.5</td>
<td>36.88</td>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>40</td>
<td>40.5</td>
<td>39.89</td>
<td>72</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>37</td>
<td>40.5</td>
<td>37.41</td>
<td>69</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>25</td>
<td>40.5</td>
<td>24.98</td>
<td>48</td>
<td>0</td>
</tr>
<tr>
<td>23</td>
<td>16</td>
<td>40.5</td>
<td>15.95</td>
<td>39</td>
<td>0</td>
</tr>
</tbody>
</table>

that the service times are distributed uniformly within 20% of the expected service time. We learnt in this class earlier that under the steady state conditions, the expected queuing delay per customer at a M|G|1 queuing system is given by a constant plus a term that varies linearly with variance or quadratically with the standard deviation. Even under dynamic conditions, we find that the expected delays seemed to vary quadratically with standard deviation in service times assuming constant expected service time. This is confirmed by fitting a quadratic curve against the data in figure 2 which provides a 99.94% fit.

Figures 2 and 3 are based on the actual demand profile for Philadelphia International Airport (PHL) on 31st January 2008 and capacity values are based on
3.3 Number of servers

Number of servers is assumed to be one. In reality, the number of servers may be equated to the number of runways. However, often these runways are not independent of each other. The arrivals and departures on close parallel runways or on intersecting runways depend on each other. At many of the congested US airports that are considered in this report, under a given runway configuration, only one runway is used for arrivals. Further, the airport capacity numbers as published by FAA are aggregated at the airport level. So, it is much more meaningful and convenient to use a single server model than a multi-server one.

3.4 Implementation details

The aforementioned simulator has been implemented in the Java programming language. The number of samples was chosen to be 100,000 i.e. 100,000 repetitions were aggregated to calculate all the statistics in this report. With increasing number repetitions, the variance of simulation results decreases and statistical significance improves. So the results become increasingly more reliable. Various aggregate statistics are obtained from the simulation. We have chosen the variance in total delays as a representative statistic of variance in simulation results. We
ran the simulator 5 times and calculated the variance of simulated total delays from those 5 runs. It must be noted that this methodology is not statistically very rigorous. However, even with such simple approach, tradeoff between run-time and reliability of results can be understood intuitively by looking at figure 4. Figure 4 shows the tradeoff between run time and variance of simulation results. On the x-axis is the log of number of repetitions. On the y-axis, the log of variance of total delay and the log of run-time in milliseconds has been plotted. As number of repetitions increase, the variance decreases at the expense of increasing run-times. We tried using 100, 1000, 10000, 100000 and 1000000 samples. Too few samples resulted in the unstable results i.e. the resulting delay values differed significantly upon running the simulator multiple times. On the other hand, if too many samples are used then it increases the runtime too much. With 100 samples, the total delay had a coefficient of variation equal to 0.364% across different simulation runs, with 1000 samples it was 0.110%, with 10,000 samples it was 0.081%. With 100,000 this number reduced to 0.017% and for 1,000,000 the number reduced all the way to 0.006%. The log of run-time increased more-or-less linearly with the log of number of repetitions. The run-time was under 5 seconds for 100, 1000 and 10,000 repetitions but increased to 47 seconds for 100,000 and to almost 8 minutes for 1,000,000 repetitions. So 100,000 samples were found to be a good trade-off between the conflicting objectives of low variance of results and manageable runtimes. Note that the variance of simulated total delay becomes most critical
when we calculate the marginal delay values due to addition of one extra arrival in certain time interval. This is because the marginal or differential values will be much more sensitive to small variations in total delay.

4 Dynamic Aspects of Queuing

Having described the internal design of the simulator, we will now proceed to investigating the validity of various concepts and relationships introduced in the class.

4.1 Delays when capacity exceeds demand

Due to the probabilistic nature of queuing system, delays do occur even when the capacity is greater than demand. A cumulative diagram fails to recognize this fact because cumulative diagrams don’t consider stochasticity in demand and service times. However, with our simulator this fact can be attested. The figure 5 shows the demand, capacity and average delays at Laguardia Airport (LGA) in New York under VFR conditions. The demand profile is the actual demand profile on 31st January 2008 at Laguardia Airport.
4.2 Impact of GDP

Ground Delay Programs (GDP) is probably the most important way of minimizing the airborne delays. When the capacity at an airport is expected to be reduced due to poor weather conditions for a certain period of time, the flights that are supposed to arrive at the airport are intentionally delayed on ground at the departure airports so that most of the delays occur in the form of ground delays and the airborne delays are kept at a minimum possible level. The effects of capacity reduction, however, persist a long time after the capacity has returned to maximum level. This effects can be seen from figure 6. The capacity profile is constructed from the actual GDP scenario at the John F. Kennedy (JFK) International Airport in New York on 6th October 2007. The capacity was reduced to VFR capacity from 11 am to 5 pm on that day. Figure 6 shows the capacity and the average delays by time interval under the normal (VFR) scenario and under the capacity reduction (GDP) scenario. It can be seen from figure 6 that even though the capacity returns to VFR level at 5 pm, the average delays do not reach the VFR levels even till midnight. This shows that the impact of a GDP can last long beyond the time when the airport returns to operating at full capacity. Also, again it can be noted that the average delay peak under both VFR and GDP scenarios occurs close to the point where the demand finally goes below capacity.
4.3 Delay peak often lags behind the demand peak

The peak demand at congested airports is often associated with demand exceeding capacity in short term. Therefore, there is a finite time interval around the demand peak where the demand exceeds capacity. After that time interval demand goes back below capacity. As long as demand exceeds capacity, the queue keeps building and once demand goes below capacity the queue starts dissipating. The average delay peak corresponds to the point where queue stops building and starts dissipating instead. Such point typically occurs later than the demand peak. Therefore the delay peak lags behind demand peak. This analysis can be substantiated using the simulator. Figure 7 shows the demand, capacity and average delays at the John F. Kennedy International Airport (JFK) in New York under VFR conditions. The demand profile is the actual demand profile on 31st January 2008.

4.4 Marginal delay cost as a function of time-of-day

When a new user is added to a queuing system, the delays costs are expected to rise. However, not all the additional delay cost is borne by the additional user. A part of the increase in delay cost is due to the delay caused to the additional user. But there is also another, and often significant, part of additional delay cost that is imposed by the additional user on rest of the users of the queuing system.
Marginal delay cost can be defined as the extra delay cost imposed on the queuing system due to the addition of an additional user during a certain time period. The part of marginal delay cost that is borne by that additional user is called as the internal (marginal) cost while the part which is borne by remaining users is called as the external (marginal) cost of delay. Micro-economic analysis of optimal congestion prices emphasizes on marginal cost pricing of resources [CP70]. Therefore, calculating the external marginal cost of delay is very important for efficient demand management. We will discuss some aspects of demand management in more details in next section. For now, let us focus on the behavior of marginal delays.

Because of additional user in a queuing system, delays can be expected to increase. This is obvious and makes intuitive sense because more the congestion more are the delays. However, what is less intuitive is the extent to which total delay increases upon an increase in demand by 1 at various time intervals in the day. As we noted earlier, the average delay to users arriving at the airport during a certain time interval depends not only on the demand and capacity during the interval, but also on the demand and capacity during previous few intervals. In other words, the average delays are also a function of the history of the queue. So for an airport with a temporary rise in demand during peak hours followed by comparatively low demand, the average delays are typically maximum near the end of the peak hours. The contrary is true for marginal delays. Marginal delays at a
certain time interval are dependent on the demand and capacity at that interval and on the demand and capacity in subsequent intervals. Therefore marginal delays are high near the beginning of period of large demand. Both these effects can be tested using the simulator. Figure 8 indicates demand profile throughout a day at Laguardia Airport (LGA) New York under IFR conditions when the capacity is 35.5 arrivals/hr. Also shown in the same graph are the patterns of average and marginal delays during the same day. Because LGA happens to be a slot controlled airport, there are no significant peaking patterns observed in the demand profile. The demand reaches close to the VFR capacity at around 7 am and stays consistently high until around 10 pm. So the average delay peaks between 9 to 10 pm, however the marginal delay peaks much earlier than that.

Also it is interesting to note that at the peak, the marginal delay imposed by additional user between 12 noon to 1 pm is 10.54 hours. Out of this, the internal cost to the user is approximately 12.65 minutes of delay as seen from the average delay value during this time interval. So the remaining 10.33 hours of delays are borne by the remaining users because of introduction of an additional user. Air Transport Association [ATA08] has estimated the average direct aircraft operating cost per block minute during 2008 to be $74.10. According to this value, the marginal external delay cost is $45,907. This number can be put in the right context by comparing it with the average ticket revenue generated per flight into the LGA airport during this time interval which was approximately $12,049. So a quick marginal cost calculation using the simulator results shows the enormity of inefficiencies in the system.

This discussion on the importance of marginal delay costs sets the stage nicely for the discussion of demand management strategies in the next section of this paper.

5 Implications for Demand Management

So far we have discussed the internal design of the simulator and have drawn several insights into the dynamic relationship between demand, capacity and delays. Let us now proceed to demonstrating how this simulator can be used to test the impacts of different congestion reduction methodologies.

The capacity of system of runways is typically the most restrictive element at major airport and it is the predominant cause of most extreme instances of delays ([BBNO07]). Airport congestion occurs because the demand for airport resources comes close to capacity or exceeds capacity. So the two general direction of congestion mitigation efforts are along the lines of measures to increase the capacity and measures to decrease the demand. Increasing the runway capacity involves long
term planning and investment. On the other hand, demand management strategies can be implemented over a medium term time horizon. The two broad categories of demand management methods include the administrative controls and market-based techniques. Market based techniques primarily include congestion pricing and slot auctions or some combination thereof. At most of the European airports and at few congested US airports, some kind of administrative slot auctions are in place. A few congested airports such as London Heathrow Airport (LHR) have recently implemented some limited flavor of congestion based variation in slot fees for peak hours. However, pure congestion pricing or slot auctions have not yet been implemented at any of the airports.

Ausubel and Crampton ([AC05]) have compared the advantages and disadvantages of slot auction and congestion pricing. On one hand, slot auctions provide better stability for long term leases, and predictable and low congestion levels, but do not provide the airlines with enough flexibility in scheduling. On the other hand, congestion pricing offers airlines the flexibility to change schedules quickly but leads to less predictability of airport congestion levels. Under congestion pricing schemes, the slot prices are fixed by the administrator and the demand for slots during each time period is determined by the market. In slot auctions, the quantity of slots to be auctioned off is fixed but the market determines the slot prices. Under the administrative slot controls, a fixed number of slots are allocated to different air carriers based on various rules such as historical precedent.
Slot auctions and administrative controls, though placed at two extremes of the spectrum in terms of economic efficiency, have certain similarities in terms of delay impact. In both cases, the total number of slots to be allocated is determined by the administrator and later are allocated to different carriers. So the delay results can be simulated independent of which method is used for allocating the slots to individual carriers as long as the total number of slots to be allocated per time interval is known. We will call these two demand management methods as quantity based demand management. On the other hand, congestion pricing is obviously a price based demand management method where the price decision eventually determines the total number of slots that get allocated. In fact, microeconomic theory dictates that the slot prices should reflect the external marginal cost of delays imposed by the incremental user of the airport facility. Therefore, the quantity based and price based demand management methods require very different approaches towards delay simulation. Next subsection focuses on the quantity based methods while the subsequent one deals with the price based methods.

5.1 Quantity based demand management methods

As described earlier, in quantity based methods for demand management, the total number of slots to be allocated to all the carriers is fixed, which means that the number of operations i.e. the number of arrivals and departures per time period is determined first. The actual distribution of slots among different carriers is carried out in the next step. The expected delay and other parameters of distribution of delays under specific weather conditions are functions of the actual number of slots allocated to all carriers and are more or less independent of the actual carriers to which the slots get allocated in second step. Therefore, a decision has to be made on how many total slots should be allocated in each time period, based on the delay considerations. Our delay simulator can come in handy as a tool in answering this question.

We have chosen the Laguardia Airport (LGA) at New York as a case in point here, because the airport has been slot controlled for several years now and the tendency is to allow the number of operations approximately equal to the VFR capacity at the airport. The total demand for slots at the airport is typically much higher than available slots, therefore the airport operates throughout the day in a state where the scheduled demand is very close to the VFR capacity. Therefore, rather than having peak and off-peak hours, the airport operates with a flat demand profile for most of the day, i.e. from 7 am to 10 pm. So this airport can be used as a good illustrative example of effect of demand cap.

We evaluate two different demand management decisions. In one case, similar to the prevailing conditions at LGA, the demand per period is capped at the
VFR capacity. The demand profile being used is the actual scheduled demand profile observed at LGA on 31st January 2008. In the other case, we evaluate the impact of capping the demand at the IFR capacity. In other words, under bad weather conditions (except for extremely bad weather that happens rarely), the demand will still be at most equal to the realized capacity. Under each demand management scenario we calculate the delays under various capacity conditions. Using the actual GDP data made available from Metron Aviation©, we identify 6 different GDP scenarios. We calculate the delays under these 6 scenarios for both demand management cases. The 6 scenarios are as follows:

1. VFR conditions for whole day: The most optimistic weather/wind conditions
2. A mildly bad weather day (Capacity reduction from 1600 to 2100 hrs) similar to the situation on 1/15/2008
3. A typical bad weather day (Capacity reduction from 1500 to 2300 hrs) similar to the situation on 2/25/2008
4. A very bad weather day (Capacity reduction from 1300 to 2300 hrs) similar to the situation on 3/31/2008
5. A terribly bad weather day (Capacity reduction from 1100 to 2400 hrs) similar to the situation on 9/10/2007
6. IFR conditions for whole day: The most pessimistic weather/wind conditions in terms of duration of bad weather

For each of these 6 weather scenarios we evaluate the total delays under both demand management cases using our simulator. The details are summarized in figure 9. As expected, the total delay is reduced when the demand is capped at IFR rather than the VFR capacity. Note that the total decrease in delay increases with increase in duration of bad weather. Additionally, the percentage delay reduction also increases with increase in bad weather duration. Based on the GDP data obtained from Metron Aviation©, for one full year from 4/1/2007 to 3/31/2008, the entire year can be roughly categorized into days similar to one of these 6 categories. Table 2 shows that a total annual delay reduction of 38,487,053 aircraft-seconds can be achieved for flights arriving at LGA airport by capping the demand at the IFR capacity. This corresponds to a 47% annual delay reduction. Interestingly, for the given demand profile, the number of flights per hour under existing situation is 594 and capping the demand at IFR for each period will lead to a reduction of 25 flights per day, which is a 4.2% reduction. It is astonishing to note that a mere 4.2% reduction in operations can result in almost 47% reduction in delays.
5.2 Price based demand management methods

Under the price based demand management methods we will essentially take a look at congestion pricing. In congestion pricing, the main idea is to ensure that the full cost of externalities imposed by the user of a public facility such as airport is internalized by the facility users. Economic theory suggests that the social welfare will be maximized if the congestions price is equal to the external marginal cost of delay due to an additional user. So the objective in most efficient congestion pricing is to identify such a price. One important hurdle in any computational exercise involving congestion pricing is the knowledge of demand function. Potential users of a facility such as the potential air carriers at an airport will decided whether or not to use the airport based on the total cost incurred by the user because of use of the facility. This includes the internal delay cost, congestion toll and any other fixed charges. Demand is a function of this total cost. More the cost, lesser is the number of users demanding the service at the facility. It is difficult to estimate this function. We will assume a linear demand function in this exercise. Further, we will ignore any other fixed costs. So the demand function is given by:
<table>
<thead>
<tr>
<th>Scenario</th>
<th>No. of Days</th>
<th>Delay Saving (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best</td>
<td>198</td>
<td>30,993</td>
</tr>
<tr>
<td>Mildly bad</td>
<td>35</td>
<td>121,254</td>
</tr>
<tr>
<td>Typical bad</td>
<td>30</td>
<td>160,294</td>
</tr>
<tr>
<td>Very bad</td>
<td>46</td>
<td>183,539</td>
</tr>
<tr>
<td>Terribly bad</td>
<td>57</td>
<td>260,611</td>
</tr>
<tr>
<td>Worst</td>
<td>0</td>
<td>350,997</td>
</tr>
<tr>
<td>Total</td>
<td>366</td>
<td>38,487,053</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>105,156</td>
</tr>
</tbody>
</table>

Table 2: Delay reduction under quantity based demand management

\[
D(C_T) = \max(D_0 - \alpha C_T, 0) \quad (1)
\]

\[
C_T = C_I + C_C
\]

\[
C_T = MC(D) \quad (2)
\]

Demand is a linear function of total cost \( C_T \). \( D_0 \) is the demand in the absence of any congestion tolls. Total cost \( C_T \) is the sum of internal cost, \( C_I \) and the congestion toll, \( C_C \). The total cost should be equated to the marginal delay cost which is a function of demand. It is here where the simulator can be useful in congestion pricing computation. \( MC(D) \) is the marginal delay cost to the system due to an additional user, which can be computed by using the simulator. The simulator is run twice, once with and once without the additional user and the subtraction of total delays output by the two simulation runs gives the marginal delay due to the additional user. Finally, we use the average aircraft operating cost per block minute published by the Air Transport Association study [ATA08] to calculate the marginal delay cost from the marginal delay minutes.

Thus we can calculate the demand given a congestion toll and we can calculate the congestion toll given a demand. But they depend on each other. So we need to find the equilibrium demand and congestion toll such that the equations (1) and (2) are both satisfied simultaneously. So this is an example of a Fixed Point Problem.

We can try to solve this problem using various heuristics. In general, it is difficult to claim whether or not a particular algorithm will converge to a fixed point. Rather than investigating the theoretical properties of algorithms, we will follow a trial-and-error procedure to see if any of the simple algorithms converges to a fixed point.
The data to be used for testing the computation of congestion pricing algorithms consists of the actual demand profile at the John F. Kennedy International Airport (JFK) at New York on 31st January 2008. The capacity is set to the VFR capacity. We have seen earlier in figure 7 that the demand exceeds capacity only very briefly during the afternoon peak hours at the JFK airport. Therefore, we will assume that the congestion toll to be imposed is only for arrivals between 1500 and 1600 hours and the demand for remaining time periods stays constant.

5.2.1 Algorithm 1: Alternate

In this algorithm, we will start with some combination of congestion toll and demand. We will update the congestion toll based on demand using equation (2) and then update the demand based on the congestion toll using equation (1). We will keep alternating until the two values converge or until the maximum number of iterations is exceeded. Formally, the algorithm can be described as follows:

Start:
\[ i := 0, D^0 := D_0, C_T^0 := 0, \]
\[ \text{and } C_T^0 \leftarrow MC(D^0) \]
\[ D^1 \leftarrow \max(D_0 - \alpha C_T^0, 0), \]
while \(|D^{i+1} - D^i| > \epsilon|,
\[ i \leftarrow i + 1 \]
\[ C_T^{i+1} \leftarrow MC(D^i) \]
\[ D^{i+1} \leftarrow \max(D_0 - \alpha C_T^i, 0) \]

\(\epsilon\) is a small positive number which is the tolerance of the algorithm. This algorithm is applied to the problem of computing the optimal congestion price during afternoon peak hour at JFK as explained earlier. For marginal cost computation, the delay simulator is used under the assumption of VFR capacity.

Figure 10 illustrates the convergence performance of the algorithm. The x-axis shows the marginal cost in $ and the y-axis represents the demand during the afternoon peak hour in flights/hour. As we can observe from the figure 10, the algorithm keeps oscillating back and forth and does not converge.

The behavior seems to suggest that the algorithm is responding too fast to marginal cost changes by changing the demand drastically. So we should perhaps try to curb this effect by taking moving average of the demand from iteration to iteration.
5.2.2 Algorithm 2: Alternate with moving averages

In this algorithm we try simple averaging of demand with one prior iteration demand. This algorithm is formally described as follows:

Start:
\[
i := 0, D^0 := D_0, C^0_C := 0,
\]
and \( C^0_T \leftarrow MC (D^0) \)
\[
D^1 \leftarrow \max (D_0 - \alpha C^0_T, 0),
\]
while \(|D^{i+1} - D^i| > \epsilon|\):
\[
i \leftarrow i + 1
\]
\[
C^{i+1}_T \leftarrow MC (D^i)
\]
\[
D^{i+1} \leftarrow \frac{\max (D_0 - \alpha C^i_T, 0) + D^i}{2}
\]

When applied to computation of optimal congestion price at JFK during afternoon peak hours leads to very fast convergence. Figure 11 illustrates the convergence performance of this algorithm. Again the x-axis shows the marginal cost.
Figure 11: *Alternate with moving averages* algorithm converges very fast

in $\$ and the y-axis represents the demand during the afternoon peak hour in flights/hour. The algorithm practically converged within 11 iterations within a small tolerance value. The fixed point corresponds to optimal congestion price of $11,815 and optimal demand of 42 flights/hour.

### 6 Summary of Findings and Key Takeaways

Airport delay mitigation is a broad topic that requires understanding and appreciation of mathematical, computational, economic and policy aspects of the problem. Delay simulation is one critical component for understanding the interaction of queuing delays with demand and capacity. In this paper, we have developed a simulator for arrival delays at an airport using a simple queuing model. The simulation methodology used in this work, though not statistically most rigorous, provides insight into various interesting aspects of the problem and motivates directions for work in this fascinating topic.

Some of the important takeaways from the process of coding and testing the simulator were as follows:

- A simple simulator that goes one step beyond a cumulative diagram, by introducing some stochastic flavor provides a lot of intuition about the dynamic behavior of queues.
• While deciding among the various alternative choices of parameters, the important metrics that must be considered are stability of simulator performance and tractability of run times. There is always going to be a tradeoff between the two, and the choice should depend on the purpose for which the simulator is supposed to be used.

• There is very little extra effort necessary to incorporate very complex distributions. However, while selecting the underlying distributions of demand and service time, it is necessary to base the decision on the realism of results.

Next we focused on testing and verifying important concepts about dynamic queue behavior. The main takeaways were as follows:

• Several properties of dynamic queues are consistent with those under steady state, but there are some critical differences.

• While the average delays depend on history of the queuing system, the marginal delays depend on the demand and capacity in the future.

• Delays do occur even when capacity is considerably more than demand. Variability in demand and service should not be ignored.

A simulator such as this one can be used effectively to assess the impact of demand management strategies or any other delay reduction strategies that involve modifying the demand distribution and/or service times. In the context of the two types of demand management strategies that we tested, the main conclusions were as follows:

• It is comparatively straightforward to evaluate the delay impact of demand management strategies that control the actual quantity of resource to be allocated, than evaluating the impact of those which control the quantity implicitly by controlling the price.

• A small reduction in demand at crucial periods during the day can create a surprisingly high percentage reduction in the delays. On the flip side, there are several issues that may arise in actually allocating these resources efficiently among the different users. Although beyond the scope of this paper, it is an important question that needs to be answered before implementing any such strategy.

• Price based demand management techniques involve solving an interesting and computationally challenging problem to come up with the equilibrium
prices. The congestion pricing case study considered in this paper makes several simplifying assumptions especially about the demand behavior, but succeeds in motivating important challenges that must be answered for actual implementation of any congestion pricing based strategy.

References


