Sensitivity of National Airspace System Performance to Disturbances: Modeling, Identification from Data, and Use in Planning

Yan Wan* Sandip Roy*

August 11, 2008

Abstract

We study the sensitivity of traffic flow management (TFM) performance in the United States National Airspace System (NAS) to disturbances, such as weather-driven capacity/flow variations and gradual changes in route usage. We make the argument that these sensitivities can be roughly computed using queueing models for flow-management actions, and so postulate that performance becomes much more sensitive to disturbance in congested airspace. Next, historical data on the sensitivity of TFM performance to weather and other uncertainties is used to support the postulate of increasing sensitivity with increased congestion. Finally, we put forth the idea that performance sensitivity information can aid in planning TFM (e.g., planning airspace reconfiguration or aircraft routing), by showing that optimally- or well-designed queue banks and queue networks have very special sensitivity structures and hence that planning actions should aim to achieve these structures.

1 Introduction

The United States National Airspace System (NAS) is continuously subject to alteration. In the short term, disturbances including convective and winter weather, runway/airport maintenance, and security-related closures lead to changes in flows and capacities. Over a longer period, traffic densities increase at disproportionate rates at different locations in the airspace, while improvements/realignments in the traffic flow management (TFM) system modify both traffic patterns and capacities. While each of these variations in flows or capacities may impact the NAS performance, it is well understood that some have much more acute impact than others. For example, Sridhar and coworker’s empirical tool for predicting delays from weather and traffic counts, the weather-impacted traffic index (WITI), demonstrates that severe weather in particular regions (the Northeast and Upper Midwest) have disproportionate effect on delays [1,2]. In the same vein, improved

*Both authors are with the School of Electrical Engineering and Computer Science, Washington State University. Both authors contributed equally to this work.
flow-management strategies at critical airports or en route locations can significantly reduce delays throughout the airspace [3].

The observed hyper-sensitivity of NAS performance to disturbances (for our purposes, capacity and flow-density changes) at certain critical congested locations suggests that TFM planning should focus on such hyper-sensitive locations. We contend that new strategies for flow management—including local strategies such as airspace flow programs (e.g., [4]) and reconfiguration (e.g., [5]), as well as radical global alterations such as use of free flight [6]—must ameliorate this hyper-sensitivity to be effective. In particular, reducing hyper-sensitivity both allows the system to better withstand disturbances, and as we shall argue in Section 4, helps reduce congestion and delays overall. To implement these TFM strategies in the most effective way, we thus need to analyze their impact on NAS performance-measure sensitivities (e.g., delay or backlog sensitivities). While planning in the air traffic flow management system already implicitly accounts for sensitivities, in the sense that locations that are perceived to be bottlenecks during e.g. inclement weather are allocated more control resources, disturbance sensitivities have not been characterized in terms of traffic parameters nor systematically used for airspace planning. The purpose of this article is to introduce the notion of disturbance sensitivity in TFM planning. We do this in two steps. 1) We give a systematic methodology for modeling the sensitivity of NAS performance to disturbances/modifications and identifying these sensitivities from data. 2) We make the argument that sensitivities can inform evaluation and design of various strategies, because sensitivities throughout the NAS assume a special structure for optimal or high-performance designs.

To delineate the contributions to TFM planning made by our study, let us briefly connect it to the existing literature. First, our study builds on recent efforts to characterize the impact of weather on traffic flows and TFM performance [1,2,7,8]. From this evaluation standpoint, our work shows that sensitivities are a means for understanding the impact of weather disturbances, and gives a causation between congestion and disturbance sensitivity. Second, we notice that there is an extensive literature on developing optimal strategies for traffic flow management, including for such diverse tasks as airspace configuration, route planning (including under inclement weather conditions), terminal and en-route flow restriction, and capacity reassignment (e.g., [3–5,9–14]). While these various optimization algorithms each help in mitigating congestion, the air traffic system is so complex and extensive that practical global strategies for flow management are difficult to evaluate, let alone optimize. This is not least because characterization of useful performance measures in the presence of weather and other uncertainties is complicated. Our effort here is not meant to supplant the optimization tools developed in the literature, but rather to show that sensitivities are useful measures that can help in testing and improving flow management strategies.

The approach that we take for estimating the sensitivity of NAS performance measures to disturbances is based on queueing theory. Queues have long been used to model numerous aspects of the NAS (including departure and arrival processes, aspects of surface operations, en route flow restriction, and long-term planning, among others) [15–17]. Here, we put forth that the impact of disturbances on NAS performance can be characterized by considering the sensitivities of queue performance measures to capacity and inflow changes. In turn, we further show that flow management should be planned to equalize scaled sensitivities at the design locations in order to optimize overall NAS sensitivity.
We motivate and validate this approach, as follows:

- In Section 2, we review the use of queueing models in air traffic flow management, and present the sensitivity analysis for the prototypical M/D/1 queues. In particular, we find the sensitivity of backlogs/delays to congestion changes, showing that increased congestion leads to much higher sensitivity to disturbances.

- In Section 3, we give evidence of the increased sensitivity to disturbances in high-congestion locations using historical data as well as relevant literature, to validate the modeling approach. This validation also clarifies how sensitivities can be inferred or compared from data.

- In Section 4, we consider planning of flow management strategies from the perspective of the sensitivity analysis. Specifically, we argue that efforts that equalize sensitivities improve NAS performance, and show how this idea can be used for such tasks as controller workload redistribution and route re-planning.

2 The Queueing Model

Queueing models have been widely used to represent various en route and terminal area management restrictions acting on air traffic flows [15–17]. An advantage of using queueing models is that they provide a systematic way to analyze traffic flow statistics and hence evaluate the performance of management strategies, in the presence of uncertainties [18]. As an example, in [17], the performance measures (e.g., average delay/backlog) of various en route TFM strategies (e.g., MIT/MINIT, Time-based Metering, and Intelligent Control) are compared assuming a typical Poisson flow. In that work, MINIT and MIT restrictions are modeled as M/D/1 queues (Poisson input, deterministic single server). Furthermore, TFM actions on multiple Centers or NAS-wide can be viewed as a network of queues. Very similar queueing models have been developed for arrival and departure as well as surface traffic [15, 16]. In [19], we considered the design of both en route and terminal area TFM restrictions in a multi-Center region to achieve desired performance. By capturing the key features of the detailed queueing model in terms of flow statistics, we came up with more abstracted models (e.g., saturation model, stochastic linear model, and algebraic linear model), and by using these simplified models, we posed the NAS-wide TFM restriction design problem as a tractable constrained optimization problem. The artical [19] is especially important to our current development, since it shows that the NAS is well-represented as a network of capacitated queues.

In this article, we use the idea that the NAS can be viewed as a network of queues to inform longer-range planning of traffic flows and flow management. To begin, we use the queueing model to analyze the sensitivity of traffic delay/backlog to a disturbance, which alters congestion due to the change of either traffic flow rate or capacity. The disturbance on congestion may be either positive or negative: congestion may increase due to unexpected weather events; and it may decrease due to effective planning, e.g., airport construction, route re-planning, improved management facilities, and increased human or facility resources. Before pursuing the disturbance sensitivity analysis, let us give a description of the prototypical queueing model used for our analysis.
2.1 Model Details

Broadly, we consider a stream of air traffic flow entering/leaving a region (e.g., entering a Sector, at a fix, or arriving at an airport). A TFM action (e.g., an en route program such as an AFP or spacing for arrivals at airports) incurs backlog and delay on aircraft, while shaping the crossing flow. The TFM action can very often be modeled as a single-server queue: each incoming aircraft waits in line at the boundary, and the first one in the waiting list is served for some time (e.g., passes through the AFP region) and leaves the boundary. In particular, M/D/1 queues (deterministic single server queues) are widely used to model various TFM actions (en route, take-off, landing, taxi-in, and taxi-out). This is because the actions generally ensure the time/distance difference between two adjacent crossing aircraft, and this fixed difference can be reflected in modeling through a deterministic constant serving time, with the assumption that each aircraft has a similar speed. Because of the wide applicability of M/D/1 queueing models in modeling air traffic, we use this model for our analysis here, though similar sensitivity computations can be obtained for other queueing models.

Specifically, here we model the incoming air traffic flow as being a Poisson process with rate $\lambda$. This memoryless stochastic representation is representative of many aggregate flows in the airspace, in particular ones that are mixtures of several independent flows, see [17] and [20] for a justification. Hence, in a time interval $T$, the distribution of the number of airplanes approaching is given by the Poisson Probability Mass Function:

$$P(N = N_c) = \frac{\lambda T e^{-\lambda T}}{N_c!}, N_c = 0, 1, 2, \ldots$$

Moreover, we model a boundary action/restriction as having a (deterministic) service rate $\lambda_c$, or in other words a serving time of $\frac{1}{\lambda_c}$ (see Figure 1). This model for example could be used to represent a $\frac{1}{\lambda_c}$-MINIT restriction or an airport arrival process with AAR of $\lambda_c$. We refer to $\lambda$ and $\lambda_c$ as the inflow rate and restriction strength (or capacity) of the queue, respectively.

![Figure 1: Queueing Model](image)

2.2 Sensitivity Analysis

Based on the M/D/1 queue representation, we can find the statistics of performance measures such as backlog and delay imposed by a TFM action [17]. The mean backlog is

$$E(B) = \frac{\lambda^2}{2\lambda_c(\lambda_c - \lambda)}$$

(2)
and the mean delay is

$$E(D) = \frac{\lambda}{2\lambda_c(\lambda_c - \lambda)}.$$ (3)

Now let us study the sensitivity of delay/backlog with respect to congestion. To do so, let us define the congestion level $\rho$ as the ratio $\frac{\lambda}{\lambda_c}$. For inflow rates $\lambda$ near the restriction rate $\lambda_c$, the congestion level $\rho$ is near 1, which represents a highly congested server. Meanwhile, $\rho << 1$ implies that the region has a lot of resources (e.g., runways, airline spacing, human controllers) that are not utilized.

The congestion level at a TFM restriction is subject to change due to unexpected weather events, and due to re-organization of traffic flows/region capacities through planning. Specifically, in the face of severe weather, the restriction $\lambda_c$ is decreased to, say, $\tilde{\lambda}_c < \lambda_c$ due to the reduced capacity, and hence the congestion level is increased by $\Delta \rho = \frac{\lambda}{\lambda_c} - \frac{\lambda}{\tilde{\lambda}_c}$. Right after the weather event, the capacity returns to normal, but the inflow rate may be increased to, say, $\tilde{\lambda}$ due to the waiting delayed aircraft. Hence the congestion level is increased by $\Delta \rho = \frac{\lambda}{\lambda_c}$ from its nominal value. Similarly, congestion may increase simply because of increased traffic demand in a region. The congestion level can also be decreased through remedial strategies. For instance, re-planning aircraft routes helps to reduce the inflow rate to a boundary, say to $\tilde{\lambda}$, and hence the congestion level is changed by $\Delta \rho = -\frac{\lambda}{\lambda_c}$. Similarly, investment in airport runway expansion, reallocation of human controllers, and improved TFM decision-making schemes can increase a region’s capacity, say to $\tilde{\lambda}_c$, so that congestion level is changed by $\Delta \rho = -\frac{\lambda}{\lambda_c} + \frac{\lambda}{\tilde{\lambda}_c}$. The fact that all these different mechanisms change congestion levels indicates the importance of finding the sensitivity of backlogs/delays to congestion.

The backlog’s sensitivity to congestion level can be obtained from Equation 2 by taking the derivative of $E(B)$ with respect to $\rho$. This yields a sensitivity $S_B(\rho)$ given by

$$S_B(\rho) = \frac{\rho(2 - \rho)}{2(1 - \rho)^2}.$$ (4)

Equation 4 clearly demonstrates the nonlinearity of the sensitivity of the backlog. As seen from Figure 2, the impact of a small disturbance on congestion level becomes more critical with the increase of the congestion level. This fact suggests two important viewpoints:

- **TFM actions or airspace with high congestion levels are more sensitive to unexpected weather events.** Much more backlog can be produced due to capacity variations in these sensitive regions than in other regions. These backlogs can propagate from these sensitive regions to the network and greatly worsen TFM performance NAS-wide.

- **In terms of planning (e.g., route replanning, allocation of human controllers, facility improvement, and runway expansion),** allocating resources to the regions that have high congestion levels will reduce the backlog more effectively. Hence, these sensitive regions that have higher congestion levels are more critical in planning.

We note that a very similar analysis can be used to determine sensitivity of average queueing delays. Unlike the backlog, the queueing delay is not explicitly a function of the congestion $\rho$, and so we find it more convenient to separately compute the sensitivity to
inflow and restriction rate changes. Specifically, the sensitivity to inflow rate changes is captured by $S_D(\lambda) = \frac{1}{2(\lambda_c - \lambda)^2}$, and the sensitivity to restriction rate changes is captured by $S_D(\lambda_c) = -\lambda(2\lambda_c - \lambda) \frac{2}{(\lambda_c - \lambda)^3}$. The sensitivity analysis of delays provides insight into weather impact and planning in the same way as the analysis of backlog.

### 3 Evidence for Congestion-Dependence of NAS Sensitivities

In this section, we give evidence supporting that the sensitivity of NAS performance to disturbances varies widely with the disturbance’s location, and more particularly that queueing models predict the dependence of sensitivity on performance. This supporting evidence also clarifies how sensitivities can be identified/compared from historical data, and related to congestion measures. With the motivation that we are introducing a broad approach to planning, we pursue canonical examples for various disturbances (e.g., weather-based capacity changes and evolution of flow densities) and locations (terminal-area and en route). We note that the examples are pursued in varying levels of detail (in some cases, we give numerical verification of the queueing-theory predictions from historical data, while in other cases only citing relevant qualitative results), but each gives credence to the described sensitivity analysis.

#### Example 1: Terminal Area Delays due to Winter Weather

Severe weather—including convection, winter weather, stratus, and high winds—is the most significant cause of delays in the NAS [21]. Here, as an example, we study the impact of winter weather on departure delays. In particular, we characterize the sensitivity of delays at eight airports in the Northeast corridor to capacity reductions due to winter weather in December 2007 and January 2008, as well as in December 2006 and January 2007. We use the Aviation System Performance Metrics (ASPM) database to perform this comparison. In particular, from the ASPM data, we obtain historical Airport Acceptance Rates (AARs) and Airport Departure Rates (ADRs) as well as actual traffic demands, and hence capacity utilization or congestion level. We also obtain the
average arrival and departure delays during Instrument Flight Conditions (IFC) periods, which specify the inclement-weather periods at these airports. At the simplest level, we wish to verify that the sensitivity to weather disturbances grows with the nominal congestion level (the average congestion level during the three-month period). To this end, we have tabulated the number of excessive delay days—i.e., days in which the average delay is more than twice the average delay for the whole period—as well as the average congestion at each airport (Table 1). As a second comparison, we have also tabulated the fraction of arriving airport delayed more than one hour at each airport, as a measure of high sensitivity. We see that the number of excessive delay days exhibits a strong dependence on the nominal congestion, with lightly congested airports (Pittsburg, Providence) having only one or two excessive delay days and the busiest (New York’s Laguardia and John F. Kennedy airports) having 10-12 excessive delay days. Similarly, the busiest airports have a much larger fraction of highly-delayed aircraft. This tendency for the busiest airports to have excessive delays verifies the higher sensitivity of queueing systems with higher congestion levels.

Table 1: We tabulate excessive delay days (days in which the delay is twice the average for that terminal) and congestion at Airports in the Northeast and Upper Midwest, during December 2007 and January 2008. We also list the fractions of aircraft arriving at each airport that were delayed more than one hour During December 2006 and January 2007. High congestion airports are more likely to have excessive delays.

<table>
<thead>
<tr>
<th>Airport</th>
<th>% Congestion</th>
<th>Excessive Delay Days</th>
<th>High-Delay Fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIT</td>
<td>13</td>
<td>2</td>
<td>0.048</td>
</tr>
<tr>
<td>PVD</td>
<td>16</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>IAD</td>
<td>35</td>
<td>4</td>
<td>0.053</td>
</tr>
<tr>
<td>BOS</td>
<td>42</td>
<td>12</td>
<td>0.039</td>
</tr>
<tr>
<td>BWI</td>
<td>44</td>
<td>7</td>
<td>0.028</td>
</tr>
<tr>
<td>DCA</td>
<td>45</td>
<td>6</td>
<td>0.035</td>
</tr>
<tr>
<td>MDW</td>
<td>46</td>
<td>7</td>
<td>0.042</td>
</tr>
<tr>
<td>PHL</td>
<td>55</td>
<td>12</td>
<td>0.072</td>
</tr>
<tr>
<td>EWR</td>
<td>56</td>
<td>8</td>
<td>0.11</td>
</tr>
<tr>
<td>ORD</td>
<td>58</td>
<td>10</td>
<td>0.084</td>
</tr>
<tr>
<td>LGA</td>
<td>58</td>
<td>12</td>
<td>0.081</td>
</tr>
<tr>
<td>JFK</td>
<td>61</td>
<td>12</td>
<td>0.080</td>
</tr>
</tbody>
</table>

We can potentially obtain a more refined characterization of the sensitivity to winter weather, by accounting for variations in weather severity at the airports. Of note, Boston’s Logan airport (BOS) appears to have an unusually high number of excessive delay days; it is plausible that this excess is caused by a higher severity in the weather at BOS as compared to the other terminals. To check whether this is the case, we have compared the reduction in capacity (ADR/AAR) at BOS during IFC periods as compared to other terminals with similar nominal congestion (e.g., IAD). As a preliminary analysis, we have compared the variability in the capacity at BOS as compared to IAD, and find that the spread is indeed larger at BOS.
Finally, as a more detailed study, we have plotted daily average delays against congestion on that day (see Figure 3), for two airports (Washington’ Dulles Airport, IAD, which has moderate congestion; and Providence Airport, PVD, which has low congestion). Linear regression of the delay with respect to the congestion is also performed. We reach two conclusions from this detailed survey: first, it lends credence to the argument that a queuing mechanism underlies the delay sensitivity of the airports, and 2) it permits detailed comparison of each airport’s sensitivity. We indeed observe much higher sensitivity at IAD.

Example 2: Effect of Disturbances on En Route Delays

Weather events also engender en route delays, by forcing re-routing of aircraft along less optimal routes, restriction of flows using AFPs and MIT/MINIT restrictions, and ground-based flow management. As with terminal-area delays, we would expect en route delays to be more sensitive to weather in highly congested areas.

The variable sensitivity of NAS-wide total delay (i.e., the total of both airborne and terminal area delay, for all flight legs during a period), is borne out by the numerical delay-prediction tool of Sridhar and coworkers [1, 2]. In particular, this empirical tool regresses total delay in terms of a weather-coverage- and traffic-density- based measure known as the weather-impacted traffic index (WITI). What is important to us here is that WITI scores for certain critical regions (in particular, the Northeast and Upper Midwest) contribute disproportionately to the total delay in the regression. On the highest delay days, the WITI for these critical regions are the ones that are exaggerated. In other words, the regression coefficients for these terms are larger than for other WITI regressors, suggesting that the sensitivity of the NAS performance (in terms of per-aircraft delays) to disturbances in these regions exceeds that of other regions. Noting that the critical regions are the ones with highest congestion, the study of regional WITI matches with the prediction of increased sensitivity obtained through queueing models. In this sense, our study can be viewed as giving a causation for the dependency of total delay on WITI and regional WITI scores.

To give further evidence for this higher sensitivity, we compare the en route average aircraft delays for two city pairs over three months (Nov. 2007 through Jan. 2008) using ASPM data. In particular, we compare en route delays for DFW-to-IAD flights, which do not pass through the highly congested Northeast corridor, and for DFW-to-BOS flights, which do pass through the Northeast corridor. We find that the standard deviation in an aircraft’s en route delay is larger by roughly a factor of 1.5 for DFW-to-BOS flights even though the mean delays are similar, indicating the higher sensitivity to disturbances of the flights passing through the congested area.

Example 3: Sensitivity of Delays to Increased Traffic

The higher sensitivity of NAS performance measures at high-congestion locations is also reflected in the dependence of delays on long-term changes in traffic demands. Here, we study the dependence of average aircraft delay on traffic demand for nine airports with

---

1We have checked that the regressions meet the criterion of the $F$-test, to ensure that sufficient data is present.
Figure 3: The dependence of the average aircraft delay on the daily congestion level is shown for two airports. We notice that the average delay exhibits a weak dependence on the congestion level (the regression line is $E(D) = 0.276\rho + 22.2$) for the low congestion airport, PVD. Meanwhile, there is a stronger dependence ($E(D) = 1.168\rho - 9.334$) for the moderate-congestion airport, IAD.

varying congestion levels, over a span of 15 years. In particular, we have studied aircraft
arrivals into 9 terminals of varying congestion levels. For each terminal, we have regressed the annual average delay for arriving aircraft with respect to the total arrival traffic. As expected, the five congested terminals (JFK, LAX, ORD, PHL, and SFO) have a strong dependence of delay on traffic demand, while the remaining terminals (MSP, PHX, SEA, and SLC) have much weaker dependence, see Table 2.

Table 2: For nine airports, we regress the average delay incurred on the aircraft entering the airport in terms of the percent change in annual traffic demand. The slopes of the regression lines are shown in the table. We note that the five highly congested airports (LAX, SFO, ORD, PHL, JFK) have strong dependences while the less congested ones (PHX, MSP, SEA, SLC) have much weaker dependences.

<table>
<thead>
<tr>
<th>Airport</th>
<th>Regression Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>JFK</td>
<td>0.88</td>
</tr>
<tr>
<td>LAX</td>
<td>1.00</td>
</tr>
<tr>
<td>MSP</td>
<td>0.17</td>
</tr>
<tr>
<td>ORD</td>
<td>2.26</td>
</tr>
<tr>
<td>PHL</td>
<td>1.70</td>
</tr>
<tr>
<td>PHX</td>
<td>-0.59</td>
</tr>
<tr>
<td>SEA</td>
<td>0.38</td>
</tr>
<tr>
<td>SFO</td>
<td>1.19</td>
</tr>
<tr>
<td>SLC</td>
<td>0.13</td>
</tr>
</tbody>
</table>

4 Using the Sensitivity Analysis for Planning

In Sections 1-3, we have verified that NAS performance is acutely sensitive to some disturbances and much less so to others. Fundamentally, we expect that knowledge of sensitivities may help us in planning and evaluation of new TFM strategies, in that high sensitivities are indicative of large delays and also are of concern themselves. In this section, we give a preliminary study on the use of sensitivity information in planning air traffic flow management strategies. Specifically, we first show that the optimal flow management schemes for banks/networks of queues are ones that equalize sensitivities to local disturbances in a certain scaled sense. We then use this insight to inform planning of various traffic flow management strategies, including reconfiguration and controller-redistribution ones. We find it most convenient to develop these design results in two steps: first, for banks of non-interacting queues (which may for instance represent multiple airports, or en route congestion points with largely uncorrelated flows); and second, for a network of queues with routing among them.

A couple further notes about the ensuing development are needed. First, we note that our analysis is focused on optimizing total backlogs, however a very similar analysis can be used to minimize delays. Second, we stress that our methodology is not focused on providing precise numerical results on optimal strategies, but instead on informing various planning tasks through use of sensitivity information.
4.1 Planning for a Bank of Queues

Here we consider a bank of \( n \) queues, i.e. a set of \( n \) queues that are operating on their own, or in other words do not have traffic flowing between them (see Figure 4). We assume an inflow rate to queue \( i \) equal to \( \lambda_i \); when we are designing inflows, we shall assume that these flows originate from a single major flow of rate \( \lambda \). Each queue \( i \)'s restriction strength is denoted as \( \lambda_{ci} \). We define the congestion level related to queue \( i \) as \( \rho_i = \frac{\lambda_i}{\lambda_{ci}} \), hence the backlog of queue \( i \) is \( E(B_i) = \frac{\rho_i^2}{2(1-\rho_i)} \), and sensitivity is \( S_B(\rho_i) = \frac{\rho_i(2-\rho_i)}{2(1-\rho_i)^2} \) according to the development in the previous section. We give the designs that minimize the total backlog using two planning schemes: 1) reconfiguration of the flows; and 2) human-controller redistribution.

![Figure 4: Bank of queues](image)

4.1.1 Reconfiguration and Route Re-Planning

Reconfiguration—or redrawing of region boundaries to ameliorate human-controller workload concerns and resulting congestion—is an area of intense current research [5,11–13]. Although reconfiguration has been widely studied, however, a key difficulty lies in choosing measures to optimize. From a planning standpoint, our approach to reconfiguration may be valuable for reducing delays and sensitivity to adverse weather. In similar fashion, re-planning of routes may mitigate congestion and delay. From our perspective, both problems resolve to that of reconfiguring the inflows to reduce backlogs. Specifically, we consider the following optimization problem:

**Design Problem 1**  Consider a bank of \( n \) queues, as shown in Figure 4. Each queue \( i \) has a fixed restriction strength \( \lambda_{ci} \). We assign the inflow rate \( \lambda_i \) for each queue \( i \) so that the total backlog of the queues \( \sum_{i=1}^{n} E(B_i) \) is minimized, subject to the following constraints:

- \( \sum_{i=1}^{n} \lambda_i = \lambda \), where the total inflow rate \( \lambda \) is positive, and less than \( \sum_{i=1}^{n} \lambda_{ci} \).
- \( \lambda_i \geq 0 \).

We refer to the optimal inflow rate in queue \( i \) as \( \lambda_i^* \), and the corresponding congestion as \( \rho_i^* \).
We show in Theorem 1 that the sensitivities of the queues’ backlogs have a simple relationship at the optimum:

**Theorem 1** Consider Design Problem 1. The optimal inflow rates $\lambda^*_i$ satisfy the following condition: there exists a constant $C$ such that $\frac{S_B(\rho^*_i)}{\lambda^*_i} = C$ for all $i$.

**Proof 1** This result follows directly from constrained optimization using Lagrange multipliers [22]. Specifically, the Lagrangian associated with the objective function and constraints is

$$L = \sum_{i=1}^n B_i + C(\lambda - \sum_{i=1}^n \lambda_i) + u_i \lambda_i,$$

where the constants $C$ and $u_i$ are nonnegative. Taking derivatives of the Lagrangian with respect to $C$, $\lambda_i$, and $u_i$ for all $i$, we obtain:

$$\frac{S_B(\rho^*_i)}{\lambda^*_i} - C + u_i = 0 \quad \forall i$$

$$\sum_{i=1}^n \lambda^*_i = \lambda$$

$$u_i \lambda^*_i = 0 \quad \forall i$$

Assuming $u_i = 0$ and solving the first two equations in Equation 5, we obtain $\lambda^*_i > 0$ for all $i$. From convexity, this solution is the global optimum. Thus, we see that $\frac{S_B(\rho^*_i)}{\lambda^*_i} = C$ for all $i$. □

This theorem shows that for the optimal flow allocation, the ratio between sensitivity and restriction strength is common among all queues. The proof of the theorem also gives the design of the optimal flow allocation: by rearranging the conditions given in the theorem together with the first constraint in the problem formulation, we can obtain the optimal flow allocation. The details of the algebra are unimportant for our purposes here.

The optimal design presented in Theorem 1 informs planning of reconfiguration and route-selection strategies in several ways:

1. The design is based on useful measures of performance (i.e., small backlog or delay), and hence permits design and evaluation with these measures in mind. Of particular interest, for a particular plan, we can check the sensitivity to restriction-strength (capacity) ratio for each congestion point, and so decide whether the performance is near-optimal. Such an approach would be helpful e.g. in evaluating the configuration design in [5], in the case where capacities vary throughout the airspace.

2. The design result suggest a data-driven methodology for iteratively improving reconfiguration strategies. In particular, from historical data, we can obtain sensitivities of backlogs/delays on various routes, as well as the capacities of the congested points along the routes. Our design shows that inflows should be reduced through route selection or reconfiguration at locations where the sensitivity-to-capacity ratio is high.
4.1.2 Controller Redistribution

Assuming that the flow rates are fixed, we can also redistribute control resources to minimize the total backlog. For instance, for en route flow restriction, human controllers can be re-assigned to mitigate capacity concerns. The problem can be formulated as follows:

**Design Problem 2** Consider a bank of \( n \) queues, as shown in Figure 4. Each queue \( i \) has an approaching Poisson flow with fixed rate \( \lambda_i \). We assign restriction strength \( \lambda_{ci} \) to queue \( i \) for each \( i \), so that the total backlog of the queues \( \sum_{i=1}^{n} E(B_i) \) is minimized, subject to the following constraints:

- \( \sum_{i=1}^{n} \lambda_{ci} = \lambda_c \) (i.e., the total capacity resource is fixed). Here, the constant \( \lambda_c \) is greater than \( \sum_i \lambda_i \);
- \( \lambda_{ci} \geq \lambda_i \).

We denote the optimal restriction strengths (capacities) by \( \lambda^*_c \), and the corresponding congestion by \( \rho^*_i \). Theorem 2 gives the structural condition of the optimal controller allocation.

**Theorem 2** Consider Design Problem 2. The optimal restriction strengths \( \lambda^*_c \) satisfy the following condition: there exists a constant \( C \) such that 
\[
\frac{S_B(\rho^*_i)\rho^*_i}{\lambda^*_c} = C \text{ for all } i.
\]

**Proof 2** The proof is very similar to the proof of Theorem 1 and hence is omitted.

Similarly to the flow reconfiguration, for the optimal controller redistribution, the backlog sensitivities of each queue are equal in a scaled sense. To obtain the optimal controller allocation, we can solve the condition given in Theorem 2 together with the first constraint given in the problem formulation.

We note that the re-distribution result also shows how new controller resources should be assigned, in particular to reduce \( \frac{S_B(\rho_i)\rho^*_i}{\lambda^*_c} \) where this measure is large. This observation may be especially helpful for planning improvement at airports, in that airports with critical need for improvement can be identified.

4.2 Planning for an Interacting Network of Queues

Finally, we study design of inflow rates for an acyclic network of queues, with the motivation of gaining insight into route-planning in a more general way. In particular, we characterize the minimum backlog design in terms of backlog sensitivities along paths in the network. We discuss application this analysis to re-allocation of routes in the NAS, either for the purpose of enacting long-term improvements in performance or for planning re-routes for common adverse-weather or high-traffic conditions. We again stress that we do not seek to capture all the details involved in route planning, but rather give a rubrik for what high-performance routing strategies are, for the purpose of planning.
4.2.1 The Network Model

We consider a network of queues that represent flows along multiple routes between one traffic source and one destination (Figure 5). Specifically, we consider a network of $n$ restrictions or queues, labeled $1, \ldots, n$, with traffic of total flow rate $\lambda$ approaching Queue 1, and leaving the network from Queue $n$. We assume that traffic flow is permitted along the edges in a directed acyclic graph, i.e. that there are a set of ordered Queue pairs $\{i, j\}$ (where WLOG $i < j$) such that traffic flow is permitted from Queue $i$ to Queue $j$. We refer to these Queue pairs as flow edges in the network, and refer to the set of such edges $E$ as the flow edge set. Without loss of generality, we assume that there is a flow path (a path of flow edges) from Queue 1 to each other queue, and similarly that there is a flow path from each queue to Queue $n$. We find it convenient to diagram the queueing network, see Figure 5. We note that an arrow is drawn from Queue $i$ to Queue $j$ in the diagram if and only if flow is permitted between the queues.

![Figure 5: Network of queues](image)

We assume that Queue $i$ has a strength or capacity $\lambda_{Ci}$ for the traffic between this source and destination. We assume that, for any set of queues whose removal separates the network into multiple pieces, the total capacity is at least $\lambda$: this requirement is necessary to permit the entire flow to traverse the network without backlog growing in time.

Meanwhile, we assume that the traffic flow from Queue $i$ to Queue $j$ is a Poisson process of rate $\lambda_{ij}$. If there is a flow edge between Queue $i$ and Queue $j$, then the flow rate $\lambda_{ij}$ is a nonnegative constant. If there is not a flow edge between the queues, then $\lambda_{ij} = 0$.

We enforce that total flow into each queue is equal to the total flow out of the queue, i.e.

$$\sum_{j \neq 1} \lambda_{1j} = \lambda$$

$$\sum_{j \neq i} \lambda_{ij} = \sum_{j \neq i} \lambda_{ji}, \quad i = 2, \ldots, n - 1$$

$$\sum_{j \neq n} \lambda_{jn} = \lambda$$

The reader will note that we have made the simplifying assumption that the flow into each Queue $i$ is Poisson; this assumption is often appropriate in air traffic management.
applications given that mixing of flows occurs between the bottleneck queues, see [17,20] for details. We refer the reader to [19] for more accurate queueing network models, which explicitly capture the effect of smoothing at one restriction on delays/backlogs at the next. We are currently pursuing routing design for one such queueing network model, namely a network of M/M/1 queues. However, these models are not critically needed for the planning tasks pursued here, so we defer a treatment to future work.

We refer to the expected backlog at Queue $i$ by $E(B_i)$, and note that the expected backlog is given by $S_B(\rho_i) = \frac{2\rho_i(2-\rho_i)}{(1-\rho_i)^2}$, where now $\rho_i = \sum_{j \neq i} \frac{\lambda_{ij}}{\lambda_{Cj}}$. We notice that the sensitivities of $E(B_i)$ to each capacity and flow rate can be computed, as in the proofs of Theorems 1 and 2.

Holistically, we refer to the model as the traffic flow queueing network.

### 4.2.2 Flow Rate Design and its Application to Route Selection

Several design problems may be of interest for the traffic flow queueing network. Specifically, as with the design for banks of queues, both capacity selection and flow rate selection can be pursued. However, noting that the inflows to each queue are assumed to be Poisson regardless of the dynamics of upstream queues, we immediately see that the capacity design problem resolves to corresponding problem for a bank of queues, and so no further development is needed. Thus, we focus here on the problem of designing flow rates between queues, to reduce backlog.

Specifically, the design problem that we address is to select the flow rates $\lambda_{ij}$ for \(\{i, j\}\) in the flow edge set $E$, so as to minimize the total expected backlog $\sum_{i=1}^{n} E(B_i)$. Here, we note that the nonnegative rates $\lambda_{ij}$ are constrained to satisfy $\sum_{i=1}^{n} \lambda_{ij} \leq \lambda_{Cj}$. We refer to this task as the flow-rate design problem. We use the notation $\lambda^*_ij$ for the optimal flow rates, and refer to this design as the optimal flow assignment. We use the notation $\rho^*_i$ for the inflow to Queue $i$ for the optimal flow assignment.

As with the design for banks of queues, it turns out that we can learn much about the optimal flow assignment by considering the sensitivities of the backlogs to the flow rates. In particular, we find that the backlog sensitivities satisfy a set of insightful conditions, as detailed in the below theorem. Before presenting the theorem, we find it convenient to define a sensitivity notion for a path. In particular, consider a path $\{j_1, j_2, \ldots, j_q\}$ from Queue $j_1 = i$ to Queue $j_q = n$. We define the total backlog sensitivity (TBS) for the path as follows:

$$TBS = \sum_{r=2}^{q} S_B(\rho_{j_r}) \frac{1}{\lambda_{Cj_r}}.$$

That is, the TBS can be computed by finding the sensitivity of the backlog to the congestion along each path edge, scaling each sensitivity by the queue capacity, and summing over the edges. Conceptually, the TBS is an aggregate measure for the sensitivity of the backlog to changes in flows along the path.

We are now ready to present the theorem on sensitivities for the optimal solution:

**Theorem 3** A flow assignment is optimal if and only if the total backlog sensitivity (TBS) along all paths from Queue 1 to Queue $n$ are the same.

In words, a flow assignment is optimal only if the TBSs along all paths from each queue to the destination are equalized; conceptually, such an assignment achieves an extremum

15
since the backlog is insensitive any differential rearrangement of flow rates (subject to the flow conservation constraints). The fact that the optimal is in fact a minimum follows naturally from the convexity of the cost. Since the analysis is quite similar to that for a bank of queues, we omit the detailed proof. Again, we note that the optimal allocation can be obtained straightforwardly from the sensitivity conditions.

The above result on the structure of optimal flow allocations is instructive for planning of routes, either for overall improvement of NAS performance or for particular common weather scenarios. In either case, the optimization result shows the following: when multiple routes from a source to a destination are available, a good route selection is one for which the total backlog sensitivity (TBS) along each path is similar. This observation can be used for route planning, as follows: from historical data, estimates of TBSs can be obtained; in turn, the sensitivities can be used to obtain improved route selections. This approach may be useful, for instance, in splitting flows among multiple routes in high-congestion or inclement-weather scenarios, see [9] for background on probabilistic planning in these circumstances.

Let us conclude our development by pointing out a couple connections and future directions of the routing-design study:

1. The result presented here is closely connected with our ongoing efforts to design controllers and/or graphs to shape an associated dynamics (e.g., [19,23,24]). These various efforts have the common theme that we identify the structure of well-designed graphs or networks, and hence compute designs that achieve high performance. Our other studies have focused on deterministic linear network dynamics; this effort is a step toward applying such structural design strategies to queueing networks.

2. Two enrichments of the presented design strategy are especially important. First, our routing design does not yet account for nominal differences in cost (e.g., delay, fuel cost) among the various options. Such differences are often present, and so making the tradeoff between nominal-cost differences and queueing costs is important. Second, our design does not explicitly try to reduce backlog sensitivity but rather only the backlog itself (though the resulting optimum is related to the sensitivities). In situations where disturbances are common, reducing the sensitivities themselves may be important.

Acknowledgement

We thank Dr. Banavar Sridhar for several illuminating conversations about planning of NAS flows and flow management. Also, we are grateful for the support provided by the National Aeronautics and Space Administration (under Grant NNA06CN26A) and the National Science Foundation (under Grant ECS-0528882).

References


