Quantifying the Tradeoff Between Fairness and Delay in Traffic Flow Management and Planning: A Queueing-Theory Approach

Sandip Roy and Yan Wan

August 3, 2009

Abstract

We examine the tradeoff between fairness and delay in air traffic flow management and planning, using an M/D/1 queue bank model. To motivate our flow-based approach to studying fairness, we first give several illustrative examples that characterize fairness (and performance in terms of delay/cost) in the current United State National Airspace System. We then pursue analysis and design of fairness, using the queue-bank model. We find that optimal allocations of capacity resources (with respect to average-delay measures) are unfair: much larger delays are imposed on minor flows as compared to major ones at the optimum. Also, the fairness of the optimal allocation is shown to exhibit a dependence, albeit a weak one, on the congestion level. We then explore how much additional delay cost is imposed if a certain fairness level is imposed, and find that significant improvement in fairness can sometimes be achieved with moderate increase in cost. Finally, we briefly note that the fairness issue studied here is analogous with the fairness concerns that arise in bandwidth allocation problems for communication networks, and suggest that further examining these commonalities may be illuminating.

1 Introduction

The purpose of this brief article is to initiate a discussion on the trade-off between cost and fairness in collaborative air traffic flow management.

The issue of fairness arises in several aspects of air traffic flow management and planning (TFM), including in 1) equitable assignment of delays among airlines or routes and 2) allocation of limited resources to improve airport or airspace performance. Recent studies of fairness in delay assignment, in particular, have yielded promising methodologies for making the impact of flow-management strategies on multiple airlines equitable (e.g., [1–5]). However, ongoing work on designing and coordinating practical flow-management strategies throughout the National Airspace System (NAS) makes clear that effective flow management and planning are targeted (see e.g. [6]), i.e. particular locations or flows or airports require
disproportionately strong flow restrictions (or, dually, disproportionately large amounts of resources for improving performance). Thus, it seems that good flow management and planning strategies are inherently unfair, in that parts of the airspace system are disproportionately impacted (or helped) by the management/planning efforts. From another viewpoint, since optimal flow management strategies seemingly often are unfair, extra cost may be incurred in seeking for fairness in flow management. The purpose of this article is to study, in a preliminary and simplistic way, the tradeoff between cost (in terms of delays, required control resources, etc) and fairness.

Let us first briefly review a couple of the viewpoints on fairness recently introduced in the air traffic management literature. In the late 1990’s, Butler and Ball, motivated by the interest in collaborative decision-making among airlines in setting ground-delay programs (GDPs), incorporated a fairness objective into an integer programming formulation for designing ground delays at a single airport [1]. More recently, Vossen and coworkers (among others) also studied fairness in ground delay programs, in particular defining an equitable resource allocation explicitly and and using this formulation to reduce inequities among airlines due to flight exemptions [2]. Further interesting work on fairness in GDPs has focused on comparing some plausible GDP allocation strategies through simulation [14,15]. Meanwhile, Green has introduced a framework in which airlines can bank delays and use these credits to seek for reduction in future delays, thus yielding fair allocation of delays [3]; this strategy is connected with ongoing studies regarding distribution and auction of airspace system resources, e.g. [4]. The concept of fairness has also been studied at a network-wide level using an Eulerian model for flow, by Raffard and co-workers. In this work, the authors propose a distributed game-theoretic solution to achieve an equitable restriction/routing of en-route traffic [5]. We note that, in addition to fair distribution of delays among airlines, fairness among different routes, destination airports, or times of day may also be important.

Of interest to us, each of these studies seeks to reallocate management efforts to maintain fairness, either through modification of an optimization objective or through game-theoretic means. At a NAS-wide level, we would expect the performance of the design (in terms of per-aircraft delay, backlog, etc.) to be markedly different from an optimal design that does not consider fairness, especially when the practicalities of implementing flow management strategies as well as uncertainties are considered. In particular, in a network-wide scenario and when resources are limited, one might expect an optimal allocation (or any high-performance practical allocation or even an allocation developed through experience) to be targeted and hence unfair [7]: consider for instance the disproportionate use of ground delay programs or ground stops at certain airports, and of en route restrictions on certain routes. Thus, a fair management effort may be significantly more costly than the naive design, i.e. there is an additional cost (in terms of delays, backlogs, infrastructure, etc) associated with achieving fairness. To the best of our knowledge, the previous studies on fairness have not sought to characterize this tradeoff between fairness and cost when practical flow management strategies are used. In this short article, we give a preliminary quantification of the fairness vs. cost tradeoff, using an abstract but canonical queue bank model for management of many flows. We find that fair solutions indeed are costly compared to optimal ones, and conversely
that optimal solutions are unfair. We also find that the degree of unfairness is significantly
dependent on the relative strengths of the two flows considered, and weakly dependent on
the level of congestion in the system. Finally, we briefly discuss how our simple insights into
the fairness-and-performance-tradeoff can inform future designs, and can be verified.

Let us stress that the flow-based study of unfairness that we pursue is part of an ongoing
effort on our part to develop collaborative traffic flow management strategies for the United
States NAS [6,7,18]. It is our belief that design of practical collaborative flow-management
requires solutions at multiple scales, ranging simple insights for planning at a NAS-wide
level to experience-based and/or optimization-based designs at a regional and local scale.
The focus of our ongoing work, including in this article, is develop simple insights into flow
management at the NAS-wide level using aggregate flow models. The study of fairness
presented here potentially can provide an enrichment of the design insights obtained in our
previous work.

Although the study of fairness in air traffic flow management (TFM) is relatively nascent,
fair allocation of bandwidth and its effect on overall network performance has long been stud-
ied in the field of *communications* (e.g., [8]). We find it more convenient in this preliminary
effort to directly model and analyze the fairness issue in air traffic systems, but note that our
formulation can be captured within the framework in the communication literature. Thus,
we believe strongly that further study of fairness in air traffic management can both draw
on and inform the efforts in the communications literature.

The remainder of the article is organized as follows. To motivate the design of flow
management for fairness, we first undertake to illustrate fairness concerns that may arise
in current NAS operation, using historical data (Section 2). In Section 3, we describe the
queue bank model and design problem, and define a measure of fairness for it. Section 4
gives analytical and simulation results regarding the performance and fairness tradeoff, and
very briefly discusses the interplay between optimality, fairness, and current practice.

## 2 Measuring Fairness in Current NAS Operations

To motivate and guide the analytical study of fairness in traffic flow management, an eval-
uation of fairness in current NAS operations is needed. To the best of our knowledge, an
evaluation from historical data has not been pursued in the literature. Here, we illustrate
numerous fairness issues that arise in current NAS operations, as motivation for the ensu-
ing analytical study of fairness. Although this preliminary evaluation of fairness in NAS
operations is not comprehensive, it does clarify several characteristics of traffic flow that
have a bearing on fairness in flow management. The numerous illustrative examples that we
present here are developed using data from the Bureau of Transportation Statistics’ *Airline
On-Time Performance Database* [17].

Equity among airlines is of primary importance because the delay of aircraft through
NAS restrictions (whether en route or at airports) incurs significant cost. Thus, each airline
has significant motivation, as a player in the TFM decision-making process, to minimize
its delays (both in an absolute sense, and relative to the other airlines). We begin our study of fairness with a comparison of NAS-incurred delays among airlines. Specifically, we have tabulated the fraction of aircraft from 10 major airlines that are subject to NAS restrictions during the period from June 2008 through May 2009, for 1) all domestic flights in the United States NAS and 2) for all domestic flights with destination at ORD (Chicago’s O’Hare International Airport) and MDW (Chicago’s Midway Airport). These statistics (Table 1 and Table 2) show that different airlines are delayed by NAS Actions at differing frequencies, both on average throughout the United States and at a particular airport. Thus, in the most basic sense, inequities in delays caused by NAS actions are present. However, many and complex factors may play a role in causing the inequities, and these factors may or may not be amenable to or deserving of resolution. For instance, in considering the frequency of NAS delays nationwide, we immediately notice that airlines that serve markets in the Western United States—where traffic is much sparser—are less likely to be delayed (with Hawaiian airlines as an extreme example), and hence NAS delay frequencies are unequal. However, at the national scale, reduction of this inequity is not immediately feasible since flow-management objectives are not met by delaying aircraft in the Western states instead of those in the East, and so equitable solutions would likely be highly non-optimal. It is worth noting that reconfiguration of the Sector boundaries could alleviate this inequity, however such reconfiguration represents a drastic alteration of current airspace operations. We also note that it is somewhat ambiguous whether reducing this inequity in delay constitutes a fair solution financially to the airlines, whose revenues clearly depend on the markets they serve.

Interestingly, considerable inequity in the frequency of NAS delays is also observed among the flights to Chicago-area airports. As expected, the airline that flies to less-congested MDW (Southwest Airlines) rather than ORD is much less likely to incur NAS delay, as are some airlines whose flights to Chicago largely originate from the more sparsely populated Western United States. We also note that likelihood of NAS delay is somewhat smaller for United Airlines, for which ORD is a hub, than for the other major carriers; we postulate that this inequity results from the higher relative prevalence of United Airline flights at ORD at low-congestion times as compared to the other airlines. In contrast with the NAS-wide inequities, we can imagine several practical strategies for mitigating inequities among the airlines for Chicago-bound traffic: possibilities include encouraging re-scheduling of flights to achieve more equal delays, using different metering of East Coast and West Coast traffic, and re-allocating human resources among nearby en route Sectors. Our previous work on high-performance management design suggests that these modifications may actually also reduce NAS delay in total (i.e., optimize performance), since dominant routes would be less restricted upon modification. Again, however, a comprehensive discussion of whether equalizing NAS impacts constitutes fairness should be initiated prior to making such changes. Nevertheless, this example suggests that design of practical flow management actions for equity and optimality is of importance. Such design will be pursued in a preliminary way in the later sections of this article.

Although we have focused on airlines’ fairness concerns in the example above, this data
Table 1: We tabulate the impact of NAS delays on 12 major airlines. Specifically, the fraction of each airline’s U.S. domestic flights that were subject to NAS delay during Jan.-Dec. 2008 are listed.

<table>
<thead>
<tr>
<th>Airline</th>
<th>% Flights NAS-Delayed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska</td>
<td>7.5</td>
</tr>
<tr>
<td>American</td>
<td>10.2</td>
</tr>
<tr>
<td>American Eagle</td>
<td>8.9</td>
</tr>
<tr>
<td>America West</td>
<td>8.2</td>
</tr>
<tr>
<td>Delta</td>
<td>11.2</td>
</tr>
<tr>
<td>Hawaiian</td>
<td>0.2</td>
</tr>
<tr>
<td>JetBlue</td>
<td>11.6</td>
</tr>
<tr>
<td>Mesa</td>
<td>6.5</td>
</tr>
<tr>
<td>Northwest</td>
<td>10.3</td>
</tr>
<tr>
<td>Southwest</td>
<td>4.0</td>
</tr>
<tr>
<td>United</td>
<td>9.4</td>
</tr>
<tr>
<td>US Airways</td>
<td>9.9</td>
</tr>
</tbody>
</table>

Table 2: We tabulate the impact of NAS delays on the 12 major airlines’ flights to Chicago. Specifically, the fraction of each airline’s U.S. domestic flights to Chicago airports (ORD and MDW) that were subject to delay during Jan.-Dec. 2008 are listed.

<table>
<thead>
<tr>
<th>Airline</th>
<th>Airport</th>
<th>% Flights NAS-Delayed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska</td>
<td>ORD</td>
<td>18.9</td>
</tr>
<tr>
<td>American</td>
<td>ORD</td>
<td>15.9</td>
</tr>
<tr>
<td>American Eagle</td>
<td>ORD</td>
<td>12.4</td>
</tr>
<tr>
<td>America West</td>
<td>ORD</td>
<td>21.6</td>
</tr>
<tr>
<td>Delta</td>
<td>ORD</td>
<td>17.9</td>
</tr>
<tr>
<td>Hawaiian</td>
<td>ORD</td>
<td>0.9</td>
</tr>
<tr>
<td>JetBlue</td>
<td>ORD</td>
<td>21.4</td>
</tr>
<tr>
<td>Mesa</td>
<td>ORD</td>
<td>5.02</td>
</tr>
<tr>
<td>Northwest</td>
<td>ORD</td>
<td>14.0</td>
</tr>
<tr>
<td>Southwest</td>
<td>MDW</td>
<td>4.3</td>
</tr>
<tr>
<td>United</td>
<td>ORD</td>
<td>12.5</td>
</tr>
<tr>
<td>US Airways</td>
<td>ORD</td>
<td>12.5</td>
</tr>
</tbody>
</table>

also exposes numerous other fairness concerns. For instance, we can infer that NAS delays more frequently impact passengers/cargo in certain parts of the United States and at certain times of day and certain days of the week. Similarly, passengers who fly through certain
airports (e.g., ORD vs. MDW) are disproportionately subject to NAS delays. The disproportionate application of restrictions may also be viewed as unequally impacting flight crews and airport personnel, among other groups. We leave a careful study of the geographic and temporal inequities for future work.

Let us pursue one more illustrative example, which exposes some further nuances in measuring fairness and crystallizes the idea of designing flow management for multiple traffic flows with fairness in mind. This example is concerned with traffic to San Francisco International Airport (SFO) during the summer months. Traffic to SFO during the summer months is rather frequently restricted because a low stratus deck at the airport often prevents concurrent landing at the two parallel runways, thus reducing the airport arrival rate and forcing flow restriction. Typically, the arrival flow is managed through ground delay programs and ground stops (with *en route* restrictions sometimes applied). We note that several groups have considered scheduling of aircraft for ground delay programs with performance and fairness in mind, and have simulated several strategies for doing so [2,14–16]. Our interest here is to understand the fairness of current implementations with respect to origin distance, and eventually to gain simple insights into fair restrictions of flows at various distances that can help with planning and rapid implementation of restrictions.

For various reasons, the ground-based restrictions are disproportionately applied to flights from nearby airports. Thus, we might postulate that NAS delays unfairly impact traffic to SFO based on the distance of the originating airport. Here, we plot the average NAS delay imposed on aircraft to SFO during June 2008 (Figure 1), as well as the average delay per mile of travel, as a function of the distance of the originating aircraft. As comparison, we also plot the delay and delay-per-mile versus distance for flights to Los Angeles International Airport (LAX), which is much less frequently impacted by stratus (or other inclement weather) and hence arrival traffic is more rarely restricted. This example introduces some further intricacies regarding fairness and flow management design. First, we see through comparison of the delay and delay-per-mile plots that the measure of impact/cost plays a key role in whether or not a flow-management solution is deemed fair. From comparison of the delays for SFO and LAX, we can confirm that short-haul flights are much more extensively restricted when destined to SFO as compared to LAX, while longer-distance flights are similarly delayed for the two airports; this comparison speaks to the inequity of restricting shorter-distance flights. On the other hand, if absolute delay is viewed as the important cost measure, we note that the restriction of flow to SFO is rather fair with respect to distance: while short-haul flights are restricted due to the arrival-rate restriction at SFO, many longer-distance flights originate from the congested Eastern United States and hence are subject to other TFM restrictions. Finally, we highlight this example as a setting where simple insights into design for fairness and optimality may assist in planning and rapid implementation of management actions (specifically, with respect to the airports chosen for ground delay). We also note that the dependence of delays on the route to the SFO is very worthwhile to study. Unfortunately, we have not had time to incorporate results in this direction, but expect to do so in an extended version of this work.
Figure 1: The dependence of NAS-imposed delay (upper plot), and delay-per-flight-mile (lower plot) on the distance of the originating airport is shown, for flights destined to SFO and LAX during June 2008. Short-haul flights to SFO are often delayed during stratus events, explaining the significant difference between SFO-bound and LAX-bound traffic. The example also highlights that the measure used to quantify fairness plays a significant role in our assessment of a flow-management action as fair or not.

The examples above illustrate the intricacies inherent to modeling and designing for fairness (equity). While clearly many details must be considered in developing equitable flow
management, several features emerge as being of core importance in designing for fairness at a NAS-wide level: 1) appropriate definition of fairness measures; 2) the importance of geography, flow-network topology, and temporal variation in flow; 3) the need for implementable designs, i.e. ones that use existing flow-management capabilities/resources and account for uncertainties; and 4) characterization of fairness-optimality tradeoffs (and comparison of fair/optimal designs with current practice) since good designs are targeted. We believe that an aggregate-flow-based design methodology is well-suited to accounts for these features of fairness at a NAS-wide level. In the remainder of the paper, we pursue a queueing-theory methodology for such design and evaluation.

3 The Model, Design Problem, and Fairness Measure

At an aggregate level, flow-management design for optimal performance and fairness can be viewed as follows: limited capacity resources (which may represent physical capacities such as runway flow rates, or human resources) must be assigned to multiple air traffic flows, so as to minimize the average delay per aircraft and also equitably distribute delays among the flows. Several design problems fit this paradigm, including human-controller (re)allocation and certain flow redistribution efforts, as well as planning efforts for airspace/airport improvement with limited resources. We do not focus on the specifics of any one of these problems here, but instead seek a simple computation of the performance vs. fairness tradeoff that gives some insight into these various problems that can assist in planning of flow management. To this end, we use a M/D/1 queue bank model for approximating delays, and so pose the capacity assignment problem and define the fairness measure. The M/D/1 queue model has been described in detail in previous work, and we advanced the queue bank model and capacity design problem in our previous study of planning using sensitivity data [7]. We review the queue and queue bank models in Sections 3.1 and 3.2, and then introduce the fairness measure in Section 3.3. We stress that the queue-bank design pursued here is just a step toward a comprehensive analytical study of fairness, which must account for networked flows and multiple capacitated resources; we have considered these more-general problems in our previous studies of coordinated flow management (without consideration of fairness) [6,7], and plan to adapt this study of fairness to the more general cases in future work.

3.1 The M/D/1 Queue Model

Let us review our formulation of the M/D/1 queue model from [7].

Queueing models have been widely used to represent various en route and terminal area management restrictions acting on air traffic flows [9–11]. 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 [12]. As an example, in [9], 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 [10, 11]. In [6], we considered the design of both en route and terminal area TFM restrictions in a multi-Center region to achieve desired performance by abstracting from queueing network models. In [7], we used the idea that the NAS can be viewed as a network of queues to inform longer-range planning of traffic flows and flow management based on simply-obtained historical data.

Here is the model for a single queue that we consider. 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 MIT implemented by human controllers, spacing for arrivals at airports, or a ground-delay program) incurs backlog and delay on aircraft. The TFM action can very often be conceptualized 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. 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 fairness studies can be pursued 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 [9] and [13] for a justification. For this model, the distribution of the number of airplanes approaching the restriction in a time-interval \( T \) is given by the Poisson Probability Mass Function:

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

We note that, in practice, an incoming flow is better modeled as a time-varying Poisson processes with time-varying rate \( \lambda(t) \), since traffic densities change markedly with time (reaching maxima in morning and evening in many locales). However, typically the rate varies slowly compared to the inter-aircraft times. We here focus on performance over short intervals, and hence assume a constant rate. (It is worth noting that we are interested in both busy or sparse intervals, and the comparison of the two.)

Moreover, we model a boundary action/restriction as having a (deterministic) service rate \( \lambda_c \geq \lambda \), or in other words a serving time of \( \frac{1}{\lambda_c} \) (see Figure 2). We refer to \( \lambda \) and \( \lambda_c \) as the inflow rate and capacity of the queue, respectively. In our formulation, the capacity is a designable quantity, for instance it can be increased by assigning more human controllers to a flow, improving runway function, or selecting between arrival and departure usage of a runway. We also find it convenient to define the congestion of the queue as the fraction of the capacity used, or the ratio \( \frac{\lambda}{\lambda_c} \), as is classical in the queueing literature. We note that the
congestion can be designed through design of the capacity.

![Figure 2: Queueing Model](image)

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 [9]. Of interest to us here, it is classically known that the mean delay per aircraft is

\[ E[D] = \frac{\lambda}{2\lambda_c(\lambda_c - \lambda)}. \]  

(2)

In our previous work [7], we also characterized the sensitivity of the delay with respect to the inflow rate and capacities, and argued that the sensitivities can be easily obtained from historical data. These analyses are not directly of use in our development here, and so we omit the details.

### 3.2 The Queue Bank Model and Design Problem

We study the allocation of limited capacity resources among multiple non-interacting flow-restriction points, so as to minimize the average delay per aircraft. This problem can be posed as that of assigning a fixed total capacity to a bank of queues with different inflow rates, to minimize average per-aircraft delay. We introduced a very similar queue-bank design problem (with backlog rather than delay as the cost metric) in [7]; below is the design problem with a delay cost metric.

Formally, consider a bank of \( n \) M/D/1 queues, as shown in Figure 3. Each queue \( i \) has an approaching Poisson flow with fixed rate \( \lambda_i \). We seek to assign restriction strengths \( \lambda_{ci} \) to each queue \( i \), so as to minimize the average aircraft delay \( \sum_{i=1}^{n} \lambda_i E[D_i] \), 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 capacities by \( \lambda_{ci}^* \), the corresponding congestions by \( \rho_i^* \), and the corresponding average delays for each queue by \( E[D_i^*] \). We refer to this design problem as the **optimal capacity assignment** problem.
Figure 3: A bank of queues is used to represent non-interacting air traffic flows that are subject to flow (capacity) constraints. We seek to distribute a limited capacity resource among the queues to minimize delays.

We note that the above capacity-assignment problem is representative of several typical traffic flow management/planning tasks, including 1) allocation of capacities to multiple routes entering a congested airspace during an inclement-weather event; 2) redistribution of human-controller effort among a set of flows, or at various times of day (which have different flow densities); 3) planning of new infrastructure for augmenting capacities under financial constraint. This article begins to study fairness issues for the queue bank model, and for the optimal capacity assignment in particular.

We also note that the capacity-assignment problem for the queue-bank model fits the framework for bandwidth allocation in communication networks, with the capacities being equivalent with bandwidth for each data stream and the delays being equivalent with the utility functions [8]. We also recall that capacity design for interacting traffic flows can be studied, using queueing-network rather than queue-bank models [7]. We defer the fairness analysis in the network case to future work.

3.3 A Measure for Fairness

Solving the optimal capacity assignment problem yields resource allocations that minimize average delay per aircraft but achieve different average delays for different flows. When these flows represent different economic entities, e.g. different airlines or passengers from different municipalities, these differences in delays may be viewed as unfair. Our previous work suggests that optimal or even good capacity allocations in queue banks and queue networks tend to be targeted, so such unfair delay distributions are common [6,7]. To systematically study fairness for the queue bank model (and specifically for the optimal capacity allocation), we find it convenient to define a measure of fairness. For the purpose of this article, we view a design as fair if the delays are (nearly) identical and unfair if there is a large scale factor between delays. Thus, we define the fairness factor of a capacity allocation as

$$ F = \frac{\min_i E[D_i]}{\max_i E[D_i]}, $$

(3)
i.e. as the ratio between the minimum and maximum expected delays among the queues. We see that $F \leq 1$, with $F = 1$ corresponding to a perfectly fair design and $F$ near 0 corresponding an unfair design. We will be interested in the fairness factor for various capacity allocations, including the optimal one. For the optimal capacity allocation, we will use the notation $F^*$ for the corresponding fairness factor.

We note that weighted fairness measures, where delays in some high-priority traffic flows are considered more costly than in the other flows or where delay per unit flight distance captures the cost, may often be of interest. We do not pursue this case further here, but note that similar analyses are possible.

4 Results

In this section, we study the fairness of optimal and non-optimal capacity allocations in the queue bank model, by characterizing the fairness factor. We first present a couple general results regarding the optimal resource allocation and corresponding values of the fairness measure. However, in this first effort, our primary focus is on exploring fairness issues through some canonical examples, and drawing interesting conclusions from these. Most of the results that we obtain either characterize the fairness of the optimal capacity allocation, or characterize the average delays of fair capacity allocations. Broadly, we find that there is indeed a tradeoff between fairness and performance (average delay).

Let us begin with a couple general results regarding the fairness of optimal capacity allocations. The first theorem compares congestions and delays for queues in the bank with different inflow rates, for the optimal resource allocation.

**Theorem 1** Consider a bank of $n$ queues, and assume WLOG that the inflow rates satisfy $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. When an optimal capacity allocation is applied to the queue bank, the congestions satisfy $\rho_1^* \leq \ldots \leq \rho_n^*$, i.e. more capacity resources are placed per unit on the queues with lower inflow rate. Nevertheless, the aircraft in the larger (higher-inflow-rate) flows are less delayed than ones in the small flows, i.e. $E[D_1] \geq E[D_2] \geq \ldots \geq E[D_n]$. Thus, the fairness factor equals $\frac{E[D_n]}{E[D_1]}$. Furthermore, each inequality is strict whenever the corresponding inflow rates are different, and hence a unity fairness factor is achieved only if all inflow rates are identical.

The proof of the this theorem builds on the sensitivity-based optimal design of queue capacities given in our previous work [7], and so we omit the details. Briefly, from the sensitivity-based characterization of the optimum, the congestion inequalities follow from a simple majorization argument, and the delay inequalities can be obtained algebraically subsequently. We kindly ask the reader to see [7] for the relevant background.

The theorem above shows that the optimal capacity allocation is unfair, with aircraft in major flows achieving lower average delays than those in minor ones. This result is not surprising, since it behooves the designer to reduce delays in the major flows to minimize...
overall delay. What is surprising is that the inequity in delays remains even though the optimal design actually assigns more capacity resource per unit (or achieves a lower congestion) for the minor flows.

We have seen that optimal designs when inflow rates are unequal are unfair. One might postulate that the degree of fairness (as reflected in the size of the fairness factor) becomes lower as inflow rates become more unequal. The following theorem formalizes this notion in the two-queue case. After the theorem, we present an example that quantifies the relationship.

**Theorem 2** Consider a bank with two queues with inflow rates $\lambda_1 > \lambda_2$. Let us consider the fairness factor of the optimal capacity allocation as a function of $\lambda_1$, assuming the total inflow rate $\lambda_1 + \lambda_2$ is kept constant. For a fixed total capacity $\lambda_c > \lambda_1 + \lambda_2$, the fairness factor at the optimum decreases as $\lambda_1$ is increased. That is, as the inflows are made more and more unequal, the delays at the optimum become more and more unfair (for the aircraft in the minor flow).

The theorem can be proved through a simple algebraic monotonicity argument regarding expected delays as the inflow rates are changed. Let us now include an example that illustrates the dependence of the fairness on the inflow rates (Figure 3). In this example, the total inflow rate of the two queues is 6, and the total capacity rate is 7. The fairness factor is plotted in terms of queue 1’s inflow rate. We see that the fairness factor decreases markedly as the two flows’ rates become more and more different, as expected. We note that similar inequities are observed for optimal capacity allocations to more than two queues, also.

![Figure 4: The fairness factor for the optimal capacity assignment in a two-queue bank is shown, as a function of the fraction of aircraft in the major flow. As the major flow becomes dominant, the optimal allocation becomes increasingly unfair to the aircraft in the minor flow.](image)

So far, we have explored the dependence of the fairness factor (at the optimum) on the magnitudes of the flow rates. It is plausible that the fairness factor also depends on
the the amount of capacity resource that is available for allocation. In particular, one might conjecture stricter capacity constraints will yield less-fair optimal allocations, because the capacity resources must be disproportionately targeted to the major flows in this case. In fact, our analyses of this case show such a positive dependence between the capacity constraint and fairness, though the dependence is rather weak. The following two-queue example illustrates the dependence. In this example, the inflow rates to the two queues are 4 and 2. We vary the total capacity between 7 and 10 and plot the fairness factor as a function of the capacity (Figure 4). We note that, although the fairness factor does not change very much with the capacities, the absolute delays are much larger for smaller capacities; thus, the lack of fairness may be especially troubling to the stakeholders in this case.

![Fairness vs. Capacity at Optimum]

Figure 5: The fairness factor for the optimal capacity assignment in a two-queue bank is shown, as a function of the available total capacity. The optimal capacity assignment becomes more fair with larger available capacity, but the dependence is weak.

In addition to characterizing the fairness of optimal capacity allocations, we have studied the fairness and performance (in terms of average delay) of non-optimal allocations in a two-queue example. This study permits us to quantify the loss in performance that results when fairness is enforced. In particular, let us consider a scenario with a major flow and minor flow (with inflow rates of 5 and 1, respectively), and a total capacity of 7. Figure 5 shows the best achievable delay assuming that a particular fairness factor is required, for fairness factors between 0.52 (which corresponds to the optimal design) and 1. We see that there indeed is a tradeoff between fairness and performance, although the dependence is fairly weak: we can significantly improve fairness while only moderately increasing delay costs.

Let us very briefly revisit the examples concerning fairness introduced in Section 2, from the point of view of the queue-bank design given in this section. At a first glance, what stands out is that the current operation of the NAS in the examples considered is neither optimal nor fair, at least not by the delay-based measures considered above: disproportionate resources are not provided to the larger flows, and they are delayed more, not less, than the smaller flows. From this viewpoint, it is interesting to note that the fair solution lies “in between” current practice and the optimal solution, requiring more resource allocation to
Figure 6: The best achievable delay through capacity assignment when a particular fairness factor is enforced is shown, for a two-queue example. Enforcing fairness does cause non-optimal performance, however the increase in delay costs is often only moderate.

the major flows than current practice—but not as much as is needed for optimality. We strongly caution, however, that these interpretations are extremely simplistic and do not account for many of the intricacies inherent to analyzing and designing flow management actions. Nevertheless, they provide a first indication of how planning of collaborative flow management can use the results of our study. We intend to pursue application of the design methodology for improved en route management and GDP planning for traffic to SFO and other congested airports, in future work.

5 Future Work

We have obtained a preliminary analysis of the tradeoff between fairness and delay in air traffic flow management and planning efforts, using a multiple-queue-bank model. Let us note our short- and long-terms plans regarding work on fairness in air traffic management:

- In the short term, we expect to refine and validate the queue-bank model for evaluating fairness, and to develop further results regarding the fairness-delay tradeoff. Specifically, we will compare model predictions with fairness/delay values obtained from detailed numerical optimization methods and/or historical data. We will also refine our analysis to include the queue-network case, and further study non-optimal designs with good fairness properties (for instance, we will characterize the level of non-optimality as a function of congestion). We will also pursue design based on a cost that incorporates both fairness and average delay, and study other measures of fairness.

- In the longer term, we expect to make explicit the connections of this research to the studies of fairness in the communications literature, and to pursue the study of fairness more broadly for complex infrastructure networks (with application in e.g. virus spread
control or power systems analysis). We believe the connection to the communications literature to be compelling because the flow-control algorithms used in this field are (by necessity) flow-based, and so capture many of the practicalities and uncertainties that also prevail in air traffic flow management and other infrastructure-network management problems. On the other hand, we stress that safety and performance requirements are much more stringent for air traffic management (and often for other infrastructures) as compared to data-communication applications, and the study of fairness must be adapted with this core distinction in mind. In studying fairness from a communication/control standpoint, we are especially interested in understanding how a network’s topological structure can be exploited to obtain high-performance yet fair designs.

References


