Can Time Buffers Lead to Delays? The Role of Operational Flexibility

Milind G. Sohoni[†] Sanjiv Erat[‡]

February 23, 2015

Abstract

How does the time buffer (slack) in an operating system affect its performance, and why? In this paper, we consider operating systems which have significant uncertainty associated with the feasible start time, thus making the actual (or available) time buffer distinct from planned (or scheduled) time buffer. The article proposes a stylized model that explicitly accounts for operational flexibility, and examines how these two time buffers (scheduled and actual) affect delays. In addition, we also evaluate the empirical content of our model by taking its predictions to the real-world data from the domestic airline industry, and examine the role of time buffers in driving operating performance (delays). Our empirical results demonstrate that, consistent with our stylized model, both scheduled and actual time buffers affect operating performance. Specifically, *smaller actual time buffers* and *larger scheduled time* buffers are associated with *greater delays*. Moreover, consistent with our model, we find that both these effects are moderated by operating flexibility. Overall, our results highlight the importance of understanding both the direct effect of time buffers and the role of resource flexibility to manage operational performance.

[†]The Indian School of Business, Gachibowli, Hyderabad 500032, India; milind_sohoni@isb.edu
[‡]The Rady School of Business, UC San Diego, La Jolla, CA 92093; serat@ucsd.edu

1 Introduction

Managing many operating systems involves building schedules (plans) that outline the time available to complete a sequence of operational[1] tasks along with the associated deadline for completion of the task. For instance, in knowledge-intensive domains such as project management, start time of an activity is planned depending on when the preceding activity completes and based on the deadline (for instance, see Kelley and Walker 1959, Berman 1964). Similarly, in made-to-order manufacturing, firms plan their production schedule based on the delivery date and when the inputs (raw materials) become available (Rajagopalan 2002).

In all such operating systems, the *scheduled time buffer* - which we define as the time from the planned start of the activity to the time of deadline - is only a tentative projection. The actual start time depends on several variables, and even with a plan that requires starting an activity at a specific given time, the actual start time is often subject to many uncertainties such as late completion of preceding activities. Hence, the available time buffer or the *actual time buffer* - which we define as the time from the actual start time of the activity to the time of deadline - might be very different from the scheduled time buffer. This article is concerned with understanding how the time buffers (scheduled and actual) influence the performance of general operating systems.

We motivate our study with a specific example from the service-oriented airline industry: Airlines develop detailed plans which lay out when a specific aircraft is scheduled to arrive at and depart from an airport. This plan determines the scheduled time buffer, the planned time (*scheduled ground-time*) that the airlines have available to conduct all the activities such as deplane arriving passengers, cleaning and refueling the craft, and boarding new passengers. But the uncertain actual arrival time of the aircraft on the day of operation implies that the actual time buffer (*actual ground-time*) available might be different from the scheduled time buffer. This might affect operating performance measured through missed deadlines

[1] We use the terms operational and operating interchangeably throughout the paper.

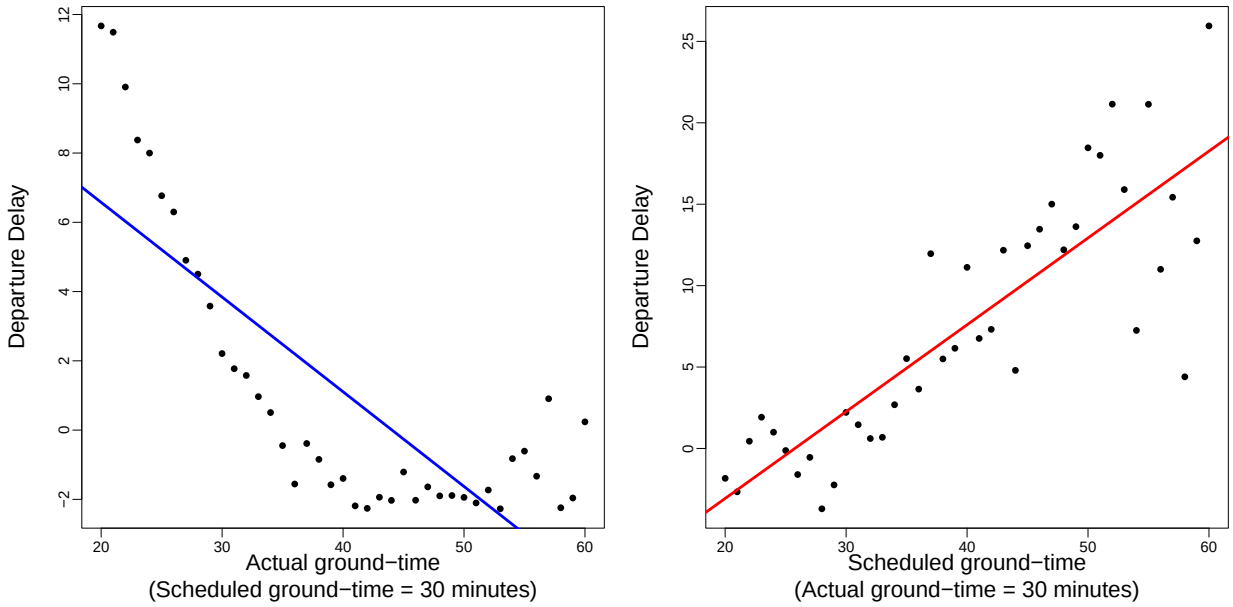


Figure 1: Departure delays for domestic US flights in November 2012.

and departure delays.

Using the publicly available data from Bureau of Transportation Statistics (BTS) website[2], we constructed the scatter plots in Figure 1, that show the fitted lines for flight departure delays as a function of actual time buffer (left) and scheduled time buffer (right). The effect of the actual time buffer on operating performance is intuitive and merits limited discussion. Specifically, one should expect that when the actual time available is lower (higher), the operating performance is likely to be poorer (better). However, the effect of the scheduled time buffer, as can be seen in the right panel of Figure 1, is more subtle. Indeed, while one may anticipate that the scheduled time buffer (after accounting for actual ground-time) should not affect operating performance, the figure in fact shows clear evidence that the departure delay is greater when scheduled time buffer is larger!

Parkinson's law and theories of procrastination offer one obvious behavioral approach to understand the "unusual" relationship between scheduled time buffers and operating performance. Specifically, such theories (perhaps in a half-serious vein) argue that work expands to

[2]http://www.transtats.bts.gov/tables.asp?DB_ID=120&DB_Name=&DB_Short_Name= (last accessed February 26, 2014).

fill the available time, and thus suggest that scheduling larger time buffers may not just fail to improve operating performance, but may in fact reduce it (for instance, see Schonberger 1981, Gutierrez and Kouvelis 1991a, Rand 2000, Goldratt 1997). While the behavioral Parkinson's law might be one of the reasons, we propose an alternate formal operational model that has a key implication that once we account for operational flexibility, the exact same relationship between scheduled buffers and operating performance will appear by virtue of firm attempting to minimize operating costs. Thus, our formal model and its results demonstrate that what might appear to be a behavioral phenomenon can just as easily emerge as the outcome from a profit maximizing (or cost minimizing) firm's rational decision-making.

Still, given that the evidence of the relationship between scheduled buffers and operating performance does not allow us to discriminate between our rational operational model and other behavioral models of procrastination (such as Parkinson's law), we also explore additional implications of our model and offer a set of theoretical propositions about how the time buffers (actual and scheduled) interact with other operational variables in determining operating performance. Specifically, only our model predicts that flexibility of an operating system moderates the degree to which time buffers influence operating performance, and that with greater flexibility the effect of both scheduled and actual time buffer is smaller.

These predictions are then tested with the airline data to compare the empirical content of our model (versus theories of procrastination). Our results bear out the key theoretical predictions of our model, and thus demonstrate that what seems to be a conventional case of Parkinson's law might in fact be the optimal behavior of an operating system. Moreover, the specific empirical results we find - the impact of buffers on performance depends on the operational flexibility - highlights the importance of understanding the extent of operational flexibility when ex-ante allocating buffers.

The rest of the article is organized as follows. In §2, we give a selective review the vast literature on planning in operating systems and time buffer management. Next, an analytical model of operational flexibility and time buffers is offered in §3, and its implications are

elucidated. §4 offers an empirical test of our key predictions. Finally, §5 offers a discussion of the practical significance of our findings, both for the case of general operating systems and for the specific case of the domestic airline industry, and concludes with some suggestions for future research. All proofs are provided in Appendix A and supplementary results are in Appendix B.

2 Literature Review

Our work broadly is related to three areas: (*i*) studies in project management (planning), and Parkinson's law and procrastination, and (*ii*) the large body of literature related to flexibility in operational systems, and (*iii*) research in airline industry that examines flight delays and its causes.

In project management, planners routinely include safety buffers above and beyond the actual time needed to complete the task so as to improve the probability of completing a task on time. However, this may result in two critical issues. First, adding such buffers could result in beginning the task as late as possible (the student syndrome or procrastination), or second, even increasing the amount and complexity of the task itself. Consequently, the addition of safety buffer may not serve its intended purpose, and the task may get delayed nevertheless. As discussed in Goldratt (1997), delays in a specific task's completion are passed on to the entire project whereas benefits from a task finishing early are rarely passed on. Goldratt (1997) proposes the *critical chain method* to address some of these problems.

The impact of introducing time buffers has also received attention in the literature related to Parkinson's law (Gutierrez and Kouvelis 1991b), and more generally, theories of procrastination, i.e., behavioral approaches to understanding how planned safety time (buffers) relate to performance. Specifically, such theories claim that work expands to fill the available buffer time, and thus suggest that scheduling larger time buffers may not achieve their intended result of reducing operational delays (Schonberger 1981, Gutierrez and Kouvelis 1991a, Rand 2000, Goldratt 1997). In similar vein, other studies consider time-inconsistent preferences

to demonstrate that people will often put off work for later (Akerlof 1991, Laibson 1997, O'Donoghue and Rabin 1999, 2001), and in some cases till the task deadline is too close resulting in task abandonment. From the psychology side, researchers have looked at slack and its effect on procrastination and found that slack (or lack of constraints) can influence the extent of procrastination (Shu and Gneezy 2010). Indeed, Ariely and Wertenbroch (2002) argue that this very behavioral trait might result in people being willing to self-impose meaningful deadlines to overcome procrastination and hence reduce delays. Zauberman and Lynch Jr (2005) make the argument that potential for future slack results in people deferring more work, and that this can lead to procrastination (and delays) when tasks have a greater slack.

Our paper is also related to the extensive work in operations management that examines the value of operational flexibility. Jordan and Graves (1995) is one of the seminal papers that demonstrates the value of flexibility in an operating system. Building on this paper, Graves and Tomlin (2003) demonstrates how to analyze the benefits of process flexibility in multi-product supply chains facing uncertain demand. Van Mieghem (1998) considers the question of investing in flexible resources as a function of costs, prices, and demand uncertainty across two products. Van Mieghem (1998) argues that, contrary to conventional wisdom, investing in flexible resources is advantageous even when demand is perfectly positively correlated across the two products. In this paper we use a notion of resource flexibility similar to that in Van Mieghem (1998), and also model our cost function similarly. One of the important contributions of our paper is that we empirically demonstrate that impact of constraints (like start time and deadlines) on operational performance depends on the flexibility, and consequently the choice of buffer availability needs to be guided by an understanding of the amount of flexibility in the system.

Lastly, since we use data from the airline industry to build and test our empirical models, our study also offers insights into delay propagation in (airline) networks and its impact on operational performance. Several papers in this area (Mayer and Sinai 2003, Deshpande and

Arikan 2012, Arikan et al. 2013) have developed models of flight delays and its antecedents. Indeed, the importance of managing delays requires airlines to routinely perturb their schedules to improve operational performance by adjusting their flight block-times and ground-times (Sohoni et al. 2011, Arikan et al. 2013). Our results, in addition to being statistically significant, show that an economically significant part of flight delays is explained purely by scheduling decisions and more specifically scheduled time buffers (which have not been previously considered in past literature). Thus, our results highlight the role of operational flexibility while scheduling and managing the ground-time slack in this industry.

3 Model Setup and Results

In this section we develop a stylized model that captures the key variable of operational flexibility to examine the impact of scheduled and actual time buffers on operating performance. Suppose that the following sequence of planning events occur at the focal firm. First, the firm determines and publishes its schedule, which comprises the scheduled start time and the deadline for a given activity. Let T denote the *scheduled time buffer*, i.e., the time from scheduled start time to the deadline. For modeling purposes we assume that the activity can be completed in K time units with minimal cost, i.e., the K represents the base-line time required to complete the activity.

Given the uncertainty associated with the operating system under consideration, we shall assume that the *actual (available) time buffer*, represented by t, may be different from T. We assume that the actual buffer time, t, is random and is distributed as $T + t_0$, where t_0 is a random variable. Once the scheduled times are published, and T and the distribution of t are known, the planner may ex-ante (i.e. during the planning stages) choose to reduce the base time K by δ_1 time units so as to complete the tasks in $K - \delta_1$ time units. For example, the planner can do so by ex-ante allocating additional resources. Similar to Van Mieghem (1998) we assume that the cost of reducing completion time (allocating additional planned resources) is linear in δ_1, i.e., $c \cdot \delta_1$.

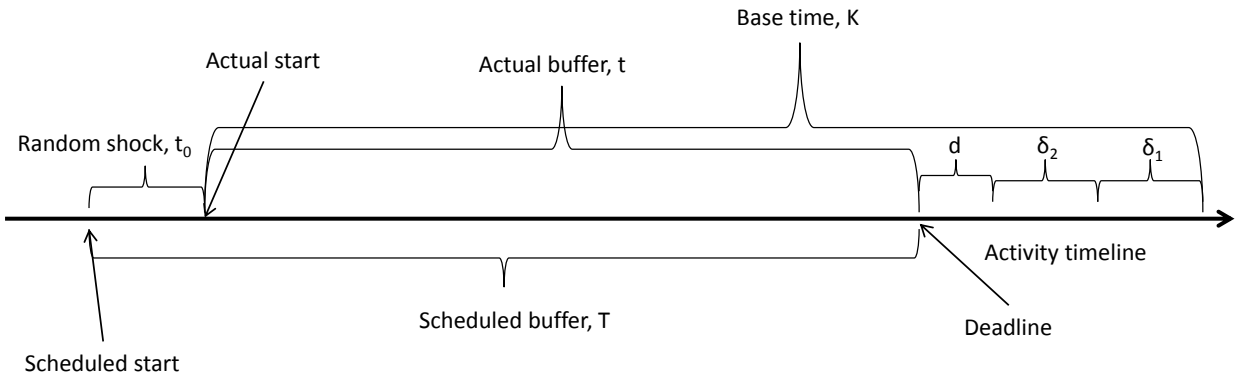


Figure 2: Sequence and timing of events and key variables.

Subsequently, during real-time operations, once t is realized, the planner may have an additional opportunity to deploy additional resources and reduce the completion time by an additional δ_2 time units. Let the cost of this real-time reduction in activity time be given by $c \cdot (1+\theta) \cdot \delta_2$. In this model, we interpret the θ parameter in terms of the amount of flexibility of the operating system, where a high θ corresponds to a larger cost, i.e, lower flexibility. Thus, the service can complete in time $K - \delta_1 - \delta_2$ for a given t. This implies, that for a given t the delay is given by $d = K - \delta_1 - \delta_2 - t$. Figure 2 graphically illustrates the sequence of events and the definitions of our key variables.

We assume that delay is costly only when it is positive (i.e., finishing the service before the deadline offers no benefits). That is, we normalize the delay cost to $\max\{d,\, 0\}$. In this setup the goal of the focal firm is to minimize the total costs; i.e., the sum of ex-ante costs $c\delta_1$ and ex-post costs $c\,(1+\theta)\,\delta_2$ and the delay costs $\max\{d,\, 0\}$. This optimization problem may be solved as follows.

In the second stage, after the actual time buffer t has been realized, the firm chooses δ_2 according to the following minimization problem:

$$\min_{\delta_2} \quad (K - \delta_1 - \delta_2 - t)^+ + c(1+\theta)\,\delta_2, \tag{1}$$

$$\text{s.t.} \qquad \delta_2 \geq 0. \tag{2}$$

We summarize the optimal values of $\delta_2^\star$ and $d^\star$ in Lemma 1.

Lemma 1. *The optimal solution to (1)-(2) is given by*

$$\delta_2^\star = \begin{cases} \max\{0,\, K-\delta_1-t\} & : c(1+\theta) < 1, \\ 0 & : c(1+\theta) \geq 1 \end{cases}, \tag{3}$$

and the corresponding delay $d^\star$ is given by $K-\delta_1-\delta_2^\star-t$.

The proof of Lemma 1 is straightforward. Hence, for brevity, we skip the details. Lemma 1 helps us roll back the second stage ex-post decision and to analyze the ex-ante expected total cost, TC.

First, consider the case when $c(1+\theta) < 1$. In this case

$$\begin{aligned} TC &= (d^\star)^+ + c\delta_1 + c(1+\theta)\delta_2^\star \\ &= c\left(\delta_1 + (1+\theta)((K-\delta_1-t)^+\right). \end{aligned} \tag{4}$$

Hence, the expected total cost, $\mathbb{E}[TC]$, is given by

$$\begin{aligned} \mathbb{E}\left[TC\right] &= c\left(\delta_1 + (1+\theta)\mathbb{E}\left[(K-\delta_1-t)^+\right]\right) \\ &= c\delta_1 + c(1+\theta)\mathbb{E}\left[(K-\delta_1-T-t_0)^+\right]. \end{aligned} \tag{5}$$

Second, consider the case when $c(1+\theta) \geq 1$. In this case

$$\begin{aligned} TC &= (d^\star)^+ + c\delta_1 + c(1+\theta)\delta_2^\star \\ &= (K-\delta_1-t)^+ + c\delta_1, \text{ and} \end{aligned} \tag{6}$$

$$\mathbb{E}\left[TC\right] = c\delta_1 + \mathbb{E}\left[(K-\delta_1-t)^+\right]. \tag{7}$$

Note that in either case, the the firm's ex-ante optimization problem that allocates the

ex-ante resources δ_1 is given by

$$\min_{\delta_1} \quad \mathbb{E}\left[TC\right], \tag{8}$$

$$\text{s.t.} \quad \delta_1 \geq 0. \tag{9}$$

To characterize the optimal solution to (8) we first show that that $\mathbb{E}\left[(K-\delta_1-T-t_0)^+\right]$ is supermodular in $(\delta_1,\, T)$ in Theorem 1.

Theorem 1. $\mathbb{E}\left[(K-\delta_1-T-t_0)^+\right]$ is supermodular in $(\delta_1,\, T)$.

The proofs of all theorems are given in Appendix A.

Notice that, using Theorem 1, it is easy to verify that the second term in ((7) and (5)), i.e., the expected total cost, is also supermodular in $(\delta_1,\, T)$. An immediate implication of Theorem 1 is that the optimal, ex-ante, time reduction $\delta_1^\star(T)$ is decreasing in T . We summarize this result in Theorem 2.

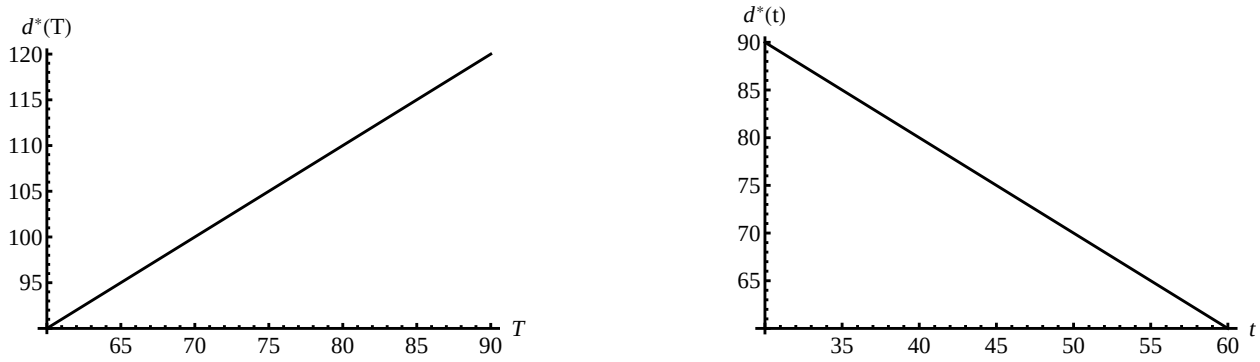
Theorem 2. *As a solution to (1) the optimal delay $d^\star$ is decreasing in t and increasing in T.*

To illustrate Theorem 2 consider the following concrete example. Let t_0 be uniformly distributed in $[t_L,\, t_H]$. For clarity, we represent both $\delta_1^\star$ and $d^\star$ as function of T. In this case we have

$$\delta_1^\star(T) = \begin{cases} K - T - \frac{c(t_H - t_L)}{1+\theta} - t_L & : \; c(1+\theta) < 1, \\ K - T + c\left(t_H - t_L\right) - t_L & : \; c(1+\theta) \geq 1, \end{cases} \tag{10}$$

and

$$d^\star(T) = \begin{cases} \min\left\{0,\, T + \frac{c(t_H - t_L)}{1+\theta} + t_L - t\right\} & : \; c(1+\theta) < 1, \\ T + c\left(t_H - t_L\right) + t_L - t & : \; c(1+\theta) \geq 1. \end{cases} \tag{11}$$



(a) $d^\star(T)$ vs. T. (b) $d^\star(t)$ vs. t.

Figure 3: $d^\star(T)$ and $d^\star(t)$ when $c = 1$, $\theta = 0.5$, and t_0 is distributed uniformly in [0, 60].

Notice that, in equations (10) and (11), $\delta_1^\star(T)$ is decreasing in T and $d^\star(T)$ is increasing in T. Additionally, $d^\star(T)$ is decreasing in t. We illustrate this in Figure 2 where $c = 1$, $\theta = 0.5$, $t_H = 60$, $t_L = 0$. The plot in Figure (3a) shows the the optimal delay, $d^\star(T)$, as T varies from 60 minutes to 90 minutes and $t = 30$ minutes. The plot in Figure (3b) shows the optimal delay as a function of t, $d^\star(t)$, as t varies from 30 to 60 minutes when $T = 60$ minutes.

Thus, our model, albeit stylized, demonstrates that the optimal delay is a decreasing function of actual time buffer, and more interestingly that the optimal delay is an increasing function of scheduled time buffer.

4 Empirical Validation and Robustness Tests

Our model of operational flexibility and delays, while parsimonious, is consistent with the actual empirical data (shown in Figure 1) that motivated this study. Still, it is to be noted that this (operational) model is not the only way to explain away the "unusual" positive correlation between scheduled slack and delays. Specifically, a more behavioral model of (time-inconsistent preferences and) procrastination also appears to be consistent with Figure 1. However, our formal model also allows us to go beyond the first order direct effects and predicts an interaction between "operational flexibility" and scheduled/actual slack; a

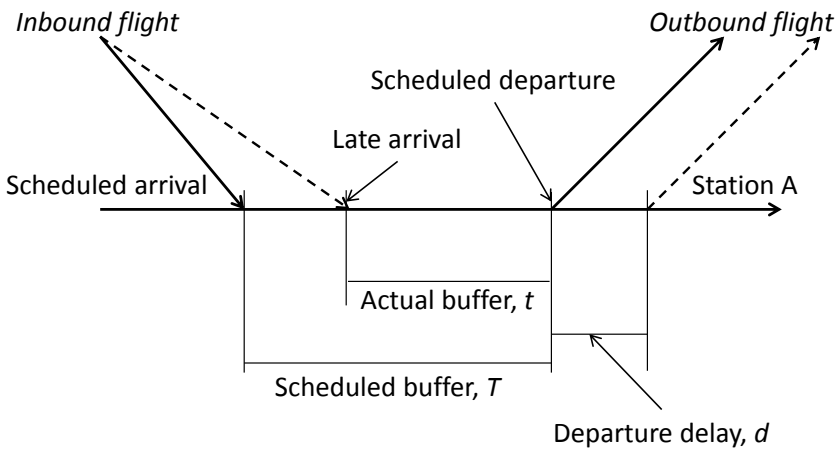


Figure 4: Flight sequence within an aircraft rotation depicting scheduled and actual ground-time buffers.

prediction that is unique to our model. The focus of this section is to offer an empirical examination of the predictive power of our model in explaining delays in operating systems versus alternate plausible explanations.

As noted earlier, while our model is meant to capture general operating systems where planned and actual slacks might differ, for the empirical test we focus on the airline industry in the US for two reasons: First, Federal Aviation Authority (FAA) requires and collects data on almost every commercial airline's operating schedule and actual arrival/departure time, which gives us access to an industry-wide large data set without having to concern ourselves with poor quality data or any selection bias. Airlines track and report five segments of the travel time for each of their flights to the FAA: (i) departure delay, (ii) taxi-out, (iii) air-time, (iv) taxi-in, and (v) arrival delay. This information is publicly available through the BTS website. Our sample is almost the entire population of civilian domestic flights operated by airlines in the US that account for at least 1% of the total domestic scheduled service passenger revenues[3]. Second, in the airline industry, the actual task that we shall empirically examine, namely servicing and getting the aircraft ready for it's next flight, is a more or less standardized task. This feature allows us to focus on the slacks and uncertainty

[3]https://www.oig.dot.gov/library-item/28632. Last accessed: December 22, 2014.

inherent in operationalizing the plans and their effect on departure delays of flights.

We begin by describing the key elements of the operating system in this industry that are relevant to our study (interested readers may refer to some examples in Holloran and Byrn 1986, Chu 2007 for some additional details). First, the airline determines and publishes its schedule. This schedule also specifies which particular aircraft, identified by their tail numbers, operate which specific flights in the network. The sequence of flights operated by a particular tail number is defined as the *aircraft rotation* (Barnhart and Cohn 2004). Figure 4 shows a sequence of inbound and outbound flights, within an aircraft rotation, at a particular station in an airline's network. The schedule also implicitly specifies the *scheduled ground-time buffer* (T), i.e. the scheduled time gap between the arrival and departure of consecutive flights in a rotation. There are several activities that must be completed during the scheduled (actual) buffer time to get the aircraft ready for departure on its next (outbound) flight. Some of these activities include de-boarding passengers, cleaning, catering, fueling, and other such activities – some that can be done simultaneously and others that must be done sequentially. Now, depending on scheduled ground-time and resource availability, the operating manager at the airport decides on how to allocate resources to complete all these tasks on time, i.e., without delaying scheduled departures.

While the planned aircraft rotation lays out the scheduled time buffer, and deadlines, the actual arrival time of the aircraft, on the day of operation, is uncertain (Arikan et al. 2013). Thus, the actual ground-time available to complete the tasks could be different from the scheduled ground-time. The operating manager has the additional ex-post ability to increase the resources and or prioritize particular tasks so as to minimize delays.

We use data from BTS website for all daily, domestic, flights flown in the US from January 2009 through December 2013. The full data set contains 31.5 million records of flights. While these records have data on the departure delay, to test our key interaction hypothesis, we also need a measure for flexibility. While operating flexibility can depend on number of different variables, for the purposes of this empirical test, we shall primarily

focus on a measure of utilization. We computed hourly flight arrivals and departures at an airport, for every airline, to estimate hourly utilization as a fraction of the maximum hourly arrivals and departures during the day. That is, we create the Utilization variable as

$$\text{Utilization (for an airline at an airport for a given day at a specific hour)} \triangleq \frac{\text{Number of arrivals and departures in that hour}}{\text{Maximum of the number of hourly hourly arrivals and departures during the day}}. \quad (12)$$

This measure of utilization, while admittedly crude, allows us to develop a proxy for operational flexibility[4] as

$$\text{Flexibility} \triangleq (1 - \text{Utilization}). \quad (13)$$

Finally, past research indicates that overall airport congestion has a significant effect on each airlines operating performance (Rosenberger et al. 2002, Lapré and Scudder 2004, Deshpande and Arikan 2012, Arikan et al. 2013). The hourly airport congestion was computed using the total number of flights, across all airlines, operating from the airport and the maximum airport arrival and departure capacity[5]. This congestion variable is, in addition to its theoretical importance, is necessary for us since our flexibility variable is likely to be correlated to utilization, and thus makes interpreting our results difficult. Hence, we shall explicitly control for congestion in our empirical tests and verify that our flexibility variable has additional predictive power above and beyond what is already explained by congestion.

Next, after creating the flexibility and congestion variables, out of of full data set of 31.5

[4]Note that most models of operating systems, including for instance queuing models, predict that higher utilization of available capacity makes any marginal increase in utilization more costly. Consistent with these models, flexibility in our analytical model (the θ parameter) captures the marginal cost in exerting effort, and thus greater utilization (which by definition) decreases our flexibility variable has the effect of increasing the cost in our stylized model.

[5]Data for each ASPM (Aviation System Performance Metrics) airport is available on FAA's Operations and Performance data website http://aspm.faa.gov. For non-ASPM airports the capacity was estimated using the maximum number of hourly flights in a given month (last accessed in January, 2014).

million, 0.5 million (1.5%) were removed either because of missing tail numbers, or because they were canceled flights. Next, using this data we built aircraft rotations[6] for each aircraft, and computed actual slack available for each flight in the rotation, t, as the difference between the scheduled departure time of the flight and the actual arrival time of the previous flight in the rotation (similar to Arikan et al. 2013). Similarly, the scheduled ground-time, T, for each flight was computed as the difference between the scheduled departure of flight and the scheduled arrival of the preceding flight in the rotation.

We removed records with missing scheduled or actual block times, or flights that were diverted (reducing the number of flights to 30.88 million). In addition, since we do not want outliers to influence our estimates, we also trimmed the data dropping the top and bottom 5% of departure delay, scheduled and actual time slack[7]. This resulted in 8.6 million less records (28%)[8]. This final cleaned data has 22.2 million observations, operated by 20 unique airlines from 332 distinct airports. Descriptive statistics of the data are given in Table 1.

	Mean	std dev
Flexibility = 1-Utilization	30.7	23.7
Congestion	65.7	20.6
Actual time slack t	156.1	201.3
Scheduled time slack T	157.3	204.5
Departure delay D	2.6	12.0

Table 1: Descriptive statistics of the airline data.

Next, we dropped all flights which had a greater than 150 minutes of slack (actual or scheduled) or less than 20 minutes of slack (actual or scheduled). These cutoffs, while somewhat arbitrary, allow us to focus was on understanding how the (scheduled and actual) slack affect delays in those interesting (and possibly more generalizable) cases where the slack is neither too small nor too large[9]. This final data that we employ in our empirical

[6] An aircraft rotation is a sequence of flights flown by a specific aircraft.

[7] These variables showed some extreme outliers. For instance, the maximum departure delay in the BTS data was 2445 minutes (40 hours)! Since this is likely the result of data error, we chose to be conservative and use only the trimmed data.

[8] Robustness checks which eliminated fewer (1%) yields similar results.

[9] Moreover, 20 minutes of ground-time is the industry standard for minimum slack, and the small number

test has 15.4 million flights, by 20 carriers, operating out of 327 airports.

Our theoretical model makes two key predictions: (i) direct effect of t and T : the delay decreases in actual time slack t and increases in scheduled time slack T, and (ii) interaction of flexibility and t, and of flexibility and T : delay is less affected by t when flexibility is high, and delay is less affected by T when flexibility is high (i.e., the interaction term for t and flexibility has a positive coefficient, and the interaction term for T and flexibility has a negative coefficient). Our empirical model, given below, controls for airline, airport, and time fixed effects and examines how the departure delay of each flight is affected by the actual and scheduled time slacks (t and T) and their interaction with flexibility.

$$\begin{aligned}\text{Delay} &= \text{Airport} + \text{Airline} + \text{Month}\times\text{Year} + \text{Flexibility} + \text{Congestion} + T + t + \text{Flexibility}\times T \\ &\quad + \text{Flexibility}\times t.\end{aligned}$$

Table 2 presents the results of our regression. As may be observed, Model 1.1 demonstrates that the effect of t and T that motivated the study is remarkably robust and is highly significant. Specifically, a 1 minute decrease in the actual slack increases the departure delay by 0.23 minute; whereas a 1 minute decrease in scheduled slack decreases the departure delay by 0.20 minutes. More interestingly, as Table 2 demonstrates, when we compare between Model 1.0 (which only includes controls that have been found in past literature to have a significant influence) and Model 1.1 (which in addition also includes actual and scheduled time slack), the R^2 goes from 4.8% to 14.5%. Note that this 200% increase in explained variance comes from just two of the operational time slack variables that we have included. Thus, it appears that the effect of actual and scheduled slacks are not merely statistically significant, but are significant from a practical perspective and crucial in determining the

of flights with less than this minimum value are possibly due to errors in data entry. In addition, the upper cut off allows us to avoid biasing our results. Specifically, many of the flights with excessive ground time are at the beginning of the rotation. Hence, we cannot infer from the data the scheduled/slack slack since even with a scheduled ground time of say 5 hours, it is unlikely that an airline would plan the start of the servicing activity this early. Still, we did conduct robustness checks by choosing a different threshold - $10 \leq t \leq 200$ and $10 \leq T \leq 200$ - and found identical results.

	Model 1.0	Model 1.1	Model 1.2
Intercept	$-3.5255^{\ddagger}(0.2216)$	$-0.7686^{\dagger}(0.2102)$	$-0.7132^{\dagger}(0.2103)$
Congestion	$0.0086^{\ddagger}(0.0002)$	$0.0074^{\ddagger}(0.0002)$	$0.0073^{\ddagger}(0.0002)$
Flexibility (=1−utilization)	$-0.0068^{\ddagger}(0.0001)$	$-0.0081^{\ddagger}(0.0001)$	$-0.0100^{\ddagger}(0.0003)$
t	-	$-0.2282^{\ddagger}(0.0002)$	$-0.2441^{\ddagger}(0.0003)$
T	-	$0.1966^{\ddagger}(0.0002)$	$0.2129^{\ddagger}(0.0003)$
Flexibility$\times t$	-	-	$0.0005^{\ddagger}(0.00001)$
Flexibility$\times T$	-	-	$-0.0006^{\ddagger}(0.00001)$
Airline, Airport, Month×Year controls	YES	YES	YES
R^2	0.048	0.145	0.145

Table 2: Regression results.

‡ $p < 0.0001$; † $p < 0.001$

delays.

Results shown in model 1.2 (shown in Table 2) are consistent with our full set of predictions - namely, the coefficient for the interaction of flexibility and actual time slack is negative and the coefficient for the interaction of flexibility and scheduled time slack is positive. Thus, delays decrease in actual time slack t, but at a smaller rate when flexibility is high; and delays increase in scheduled time slack T, but at smaller rate when flexibility is high. These interaction results allow us to rule out the simpler behavioral ("irrational") explanations of why time slack has an effect on delays and offer evidence for the empirical content of our model of operational flexibility[10].

While the key results in Table 2 are consistent with our theoretical model, the empirical measure of flexibility that we employed might conceivably bias our results. For instance, the flexibility measure being a ratio, might be prone to error if the denominator (i.e., maximum number of flights per hour operated from the given airport an airline is a small number (since in such a situation, the flexibility measure possibly fails to take too many values

[10]This is not to claim that more complex behavioral explanations can be ruled out. For instance, a model of procrastination where deferring work is also related to flexibility would be consistent with the results. Still, in our empirical model, the flexibility is the actual flexibility, and not anticipated flexibility. Thus, it appears that such a model of procrastination would also need agents whose beliefs about future flexibility need to correlated to true realized flexibility. While this is possible, we believe that our model of operational flexibility and optimal resource allocation offers a significantly simpler and consistent theory for our empirical results.

between 0 and 100%). To ensure robustness, we took a two-pronged approach. First, we re-estimated the regression equation after dropping all flights where the maximum obtained was too low (< 5) for us to reliably estimate the flexibility measure. Estimates, which are shown in Table 3 of the ancillary appendix, demonstrate clearly that our key results remain unchanged. Second, instead of using a ratio to measure flexibility, we defined an alternate measure of flexibility - namely, capacity slack - defined as

$$\begin{array}{c}\text{Flexibility (for an airline}\\ \text{at an airport for a given}\\ \text{day at a specific hour)}\end{array} \triangleq \begin{array}{c}\text{Maximum of the}\\ \text{number of hourly}\\ \text{arrivals and departures}\\ \text{during the day}\end{array} - \begin{array}{c}\text{Number of arrivals}\\ \text{and departures}\\ \text{in that hour}\end{array} . \tag{14}$$

Estimates of the effect of scheduled and actual time slack when flexibility is measured using the above definition are summarized in Table 4 of the ancillary appendix. Our results are robust and are fully consistent with theoretical model.

5 Discussion and Concluding Remarks

The current study was motivated by the fundamental issue of understanding how buffers in an operating system affect its performance. While the study of buffers in a variety of contexts and its optimal management has occupied a significant part of operations management research, there appears to be not much (empirical or theoretical) research that has examined our research question. Moreover, even in the research that has attempted to understand the impact of buffers on performance, behavioral theories and Parkinson's law feature prominently. For instance, the conventional wisdom in project management relies on the so-called "student syndrome" and theories of procrastination to argue that scheduling larger time buffer does not improve and may in fact decrease operating performance (see for example, Gutierrez and Kouvelis 1991b).

Our main contribution in this article is two fold: (i) We offer a simple and highly stylized model of how operational flexibility interacts with time buffers to determine operating per-

formance. The model is subsequently shown to imply the very same empirical patterns as implied by the behavioral theories of time buffers. Thus, our rational model offers a plausible alternative to the conventional behavioral theories that have been typically employed in past literature. (*ii*) We explicitly test our normative model against the conventional behavioral theories using data from airline industry, and we demonstrate that (at least in this context) the data is consistent with our rational model. Specifically, we show theoretically that our model implies an interesting interaction between operational flexibility and time-buffers, and that the effect of time buffers on operating performance depends on the operational flexibility. This implication, as it is not predicted by behavioral theories, allow us to do a cleaner test and results consistent with with our model are obtained.

Our results, at a general level, highlight the importance in understanding flexibility at the operational level, when choosing the optimal schedules. Specifically, a planning approach that does not account for flexibility at the operational level and tries to optimize some combination of cost/utilization and revenue/on-time-performance will end up creating schedules that are too costly and/or schedules that are poor in revenue/on-time-performance (since those who have flexibility react differently compared to those without). To improve outcome performance measure (such as on-time-performance), operating systems that lack flexibility should be constrained more (by scheduling lower time-buffers) whereas those that have significant flexibility should be constrained less. Thus, our results suggest that it is the relative weight that the firm places on the outcome and input (cost) measures, and the relative flexibility of the different parts of the operating system that determines the optimal schedule (and time buffers).

While our concern for the most part has been on developing a highly stylized model and testing it. it is still useful to note that for the specific case of airline industry that motivated our model and empirical tests, the current study offers at least two important implications. Firstly, it has long been assumed that one of the key reasons for flight delays and lower operating performance is congestion at airports and the consequent negative externality (for

example, see Rosenberger et al. 2002, Mayer and Sinai 2003). While we find that externalities imposed by congestion does play a role (Table 3), a much larger role is played by an airline's own scheduling choices. Specifically, Table 3 shows that, in our data, congestion going from 0 to 100 has a similar effect as a mere additional 5 minutes (increase) of actual slack or an additional 5 minutes (decrease) of scheduled slack! Moreover, the dramatic increase in R^2 (by over 200%) when actual and scheduled slacks are added demonstrate that the part of the variance explained by the actual operating schedules is much larger compared to congestion, airline, airport, and time taken together. Thus, for future research work that attempts to evaluate the role of congestion public policy or airline operations probably should account for an airline's own operational scheduling decisions.

Secondly, to the extent that on-time-performance is an important competitive dimension (as suggested by Deshpande and Arikan 2012), our results suggest that the differences between airlines on their on-time-performance measures may not be coming about primarily because of their routes (specific airports they fly into, or the specific airports they fly out of), or because of congestion, or because of their network structure; but that (a large part of) the differences in operating performance may be traced back to the operating schedules and more specifically to the scheduled and actual time buffers that have been allocated. Consequently, if the goal is to improve operating performance, the airlines might be better served by improvements in their planning (that accounts for operating flexibility) compared to broader strategic changes (such as changes to routes, or adding capacity at specific airports, etc).

While our theoretical model was developed with a general operating system in mind, our empirical focus in this article was on the context of flight delays and ground-time buffers. This empirical focus is mostly pragmatic (since the data are readily available). However, our findings may be applicable to wider contexts and deserves more investigation. For instance, consider a manufacturing or a project management setting. A plan, in these cases, implicitly specifies the scheduled time buffer for each activity and their completion deadlines, whereas the actual time buffer available might change depending on the uncertainty associated with

preceding activities or uncertain availability of crucial resources. Overall, our results suggest that to manage operational performance one needs to understand both the direct effect of time buffers and the role of resource flexibility in determining both the *sign* and *magnitude* of the effect of time buffers.

References

Akerlof, G. A.: 1991, Procrastination and obedience, *American Economic Review* **81**(2), 1–19.

Ariely, D. and Wertenbroch, K.: 2002, Procrastination, deadlines, and performance: Self-control by precommitment, *Psychological Science* **13**(3), 219–224.

Arikan, M., Deshpande, V. and Sohoni, M.: 2013, Building reliable air-travel infrastructure using empirical data and stochastic models of airline networks, *Operations Research* **61**(1), 45–64.

Barnhart, C. and Cohn, A.: 2004, Airline schedule planning: Accomplishments and opportunities, *Manufacturing & Service Operations Management* **6**(1), 3–22.

Berman, E. B.: 1964, Resource allocation in a pert network under continuous activity time-cost functions, *Management Science* **10**(4), 734–745.

Chu, S. C.: 2007, Generating, scheduling and rostering of shift crew-duties: Applications at the hong kong international airport, *European Journal of Operational Research* **177**(3), 1764–1778.

Deshpande, V. and Arikan, M.: 2012, The impact of airline flight schedules on flight delays, *Manufacturing & Service Operations Management* **14**(3), 423–440.

Goldratt, E. M.: 1997, *Critical chain*, North River Press Great Barrington, MA.

Graves, S. C. and Tomlin, B. T.: 2003, Process flexibility in supply chains, *Management Science* **49**(7), 907–919.

Gutierrez, G. J. and Kouvelis, P.: 1991a, Parkinson's law and its implications for project management, *Management Science* **37**(8), 990–1001.

Gutierrez, G. J. and Kouvelis, P.: 1991b, Parkinson's law and its implications for project management, *Management Science* **37**(8), 990–1001.

Holloran, T. J. and Byrn, J. E.: 1986, United airlines station manpower planning system, *Interfaces* **16**(1), 39–50.

Jordan, W. C. and Graves, S. C.: 1995, Principles on the benefits of manufacturing process flexibility, *Management Science* **41**(4), 577–594.

Kelley, J. E. and Walker, M. R.: 1959, Critical-path planning and scheduling, *Papers Presented at the December 1-3, 1959, Eastern Joint IRE-AIEE-ACM Computer Conference*, IRE-AIEE-ACM '59 (Eastern), ACM, New York, NY, USA, pp. 160–173.

Laibson, D.: 1997, Golden eggs and hyperbolic discounting, *The Quarterly Journal of Economics* **112**(2), pp. 443–477.
URL: *http://www.jstor.org/stable/2951242*

Lapré, M. A. and Scudder, G. D.: 2004, Performance improvement paths in the us airline industry: Linking trade-offs to asset frontiers, *Production and Operations Management* **13**(2), 123–134.

Mayer, C. and Sinai, T.: 2003, Why do airlines systematically schedule their flights to arrive late?, *Technical report*, Working paper.

O'Donoghue, T. and Rabin, M.: 1999, Doing it now or later, *The American Economic Review* **89**(1), pp. 103–124.
URL: *http://www.jstor.org/stable/116981*

O'Donoghue, T. and Rabin, M.: 2001, Choice and procrastination, *The Quarterly Journal of Economics* **116**(1), 121–160.

Rajagopalan, S.: 2002, Make to order or make to stock: Model and application, *Management Science* **48**(2), 241–256.

Rand, G. K.: 2000, Critical chain: the theory of constraints applied to project management, *International Journal of Project Management* **18**(3), 173 – 177.

Rosenberger, J. M., Schaefer, A. J., Goldsman, D., Johnson, E. L., Kleywegt, A. J. and Nemhauser, G. L.: 2002, A stochastic model of airline operations, *Transportation science* **36**(4), 357–377.

Schonberger, R. J.: 1981, Why projects are 'always' late: A rationale based on manual simulation of a pert/cpm network, *Interfaces* **11**(5), 66–70.

Shu, S. B. and Gneezy, A.: 2010, Procrastination of enjoyable experiences, *Journal of Marketing Research* **47**(5), 933–944.

Sohoni, M., Lee, Y.-C. and Klabjan, D.: 2011, Robust airline scheduling under block-time uncertainty, *Transportation Science* **45**(4), 451–464.

Topkis, D. M.: 1998, *Supermodularity and Complementarity*, Princeton University Press.

Van Mieghem, J. A.: 1998, Investment strategies for flexible resources, *Management Science* **44**(8), 1071–1078.

Zauberman, G. and Lynch Jr, J. G.: 2005, Resource slack and propensity to discount delayed investments of time versus money., *Journal of Experimental Psychology: General* **134**(1), 23.

A Main Appendix

Proof of Theorem 1.

Proof. It is easy to verify that $\phi(\lambda) = \mathbb{E}\left[(\lambda - t_0)^+\right]$, is a convex function in λ. To do so, let $p(x)$ denote the probability density function of t_0 and $P(x)$ denote the corresponding cumulative density function. Thus,

$$\begin{aligned} \phi(\lambda) &= \int_{-\infty}^{\infty} \max\{0,\, \lambda - x\}\, p(x)\, dx \\ &= \int_{-\infty}^{\lambda} (\lambda - x)\; p(x)\, dx. \end{aligned} \tag{1}$$

Notice that from (1), the derivative with respect to λ, $\frac{d}{d\lambda}\phi(\lambda) = P(\lambda) > 0$, and the second derivative $\frac{d^2}{d\lambda^2}\phi(\lambda) = p(\lambda) \geq 0$. The supermodularity of $\mathbb{E}\left[(K - \delta_1 - T - t_0)^+\right]$ in $(\delta_1,\, T)$ follows immediately once we substitute $\lambda = K - \delta_1 - T$ and verify that the cross-partial derivative term $\frac{\partial^2}{\partial T \partial \delta_1}\mathbb{E}\left[(K - \delta_1 - T - t_0)^+\right]$ is non-negative. □

Proof of Theorem 2.

Proof. The proof follows immediately by recalling that $d^\star = \min\left\{0,\, K - \delta_1^\star(T) - t\right\}$ and Theorem 2.8.1 in Topkis (1998). □

B Ancillary Appendix: Robustness Tests

	Model 2.0	Model 2.1	Model 2.2
Intercept	$-2.97^{\ddagger}(0.1718)$	$-0.1072\,(0.1635)$	$0.0777\,(0.1638)$
Congestion	$0.00674^{\ddagger}(0.0002)$	$0.0045^{\ddagger}(0.00023)$	$0.0044^{\ddagger}(0.00023)$
Flexibility (=1-Utilization)	$-0.0094^{\ddagger}(0.0002)$	$-0.0115^{\ddagger}(0.0002)$	$-0.01709^{\ddagger}(0.0004)$
t	-	$-0.2179^{\ddagger}(0.0002)$	$-0.2348^{\ddagger}(0.00034)$
T	-	$0.186^{\ddagger}(0.00022)$	$0.2014^{\ddagger}(0.00037)$
Flexibility$\times t$	-	-	$0.00057^{\ddagger}(0.00001)$
Flexibility$\times T$	-	-	$-0.00051^{\ddagger}(0.00001)$
Airline, Airport, Month$\times$Year controls	YES	YES	YES
R^2	0.046	0.137	0.137

Table 3: Regression results with at least 5 flights/hour/airline.
$\ddagger$ $p < 0.0001$; $\dagger$ $p < 0.001$

	Model 3.0	Model 3.1	Model 3.2
Intercept	$-4.0619^{\ddagger}(0.2215)$	$-1.3505^{\ddagger}(0.2101)$	$-1.5652^{\ddagger}(0.21)$
Congestion	$0.0145^{\ddagger}(0.0002)$	$0.0138^{\ddagger}(0.0002)$	$0.0141^{\ddagger}(0.0002)$
Flexibility (=Capacity Slack)	$0.0097^{\ddagger}(0.0004)$	$0.0077^{\ddagger}(0.0004)$	$0.0401^{\ddagger}(0.0008)$
t	-	$-0.2281^{\ddagger}(0.0002)$	$-0.2416^{\ddagger}(0.0002)$
T	-	$0.1960^{\ddagger}(0.0002)$	$0.2148^{\ddagger}(0.0002)$
Flexibility$\times t$	-	-	$0.0017^{\ddagger}(0.00002)$
Flexibility$\times T$	-	-	$-0.0023^{\ddagger}(0.00002)$
Airline, Airport, Month$\times$Year controls	YES	YES	YES
R^2	0.048	0.144	0.145

Table 4: Regression results (with Capacity Slack).
$\ddagger$ $p < 0.0001$; $\dagger$ $p < 0.001$

It is noteworthy that the results in Table (3) hold even if we consider airlines who operate at least 5 flights from any given station. The sign of the coefficient for slack is unexpectedly positive. However, note that the way we measure slack (see definition in equation 14) will be positively correlated with our measure of congestion (based on the ratio of number of flight to maximum for all airlines). Since the coefficient on congestion is negative, to find the net effect of slack, we need to do a regression without the congestion variable. This is shown in table 5. As may be verified, as capacity slack increases, the departure delay decreases.

Dependent variable: Departure Delay	
Intercept	$-0.5266\,(0.209)$
Flexibility (=Capacity Slack)	$-0.0059^{\ddagger}(0.00035)$
t	$-0.2281^{\ddagger}(0.0002)$
T	$0.1959^{\ddagger}(0.0002)$
Airline, Airport, Month×Year controls	YES
R^2	0.144

Table 5: Regression results (with Capacity Slack) and with at least 5 flights/hour/airline. ‡ $p < 0.0001$; † $p < 0.001$