Evaluating the First-Mover’s Advantage in Announcing Real-Time Delay Information

Siddharth Prakash Singh, Mohammad Delasay, Alan Scheller-Wolf
Tepper School of Business, Carnegie Mellon University, Pittsburgh PA 15213
sps1@andrew.cmu.edu, delasays@andrew.cmu.edu, awolf@andrew.cmu.edu

We propose a model of two comparable service providers competing for market share. The service providers employ different delay announcement strategies. The announcer (A) voluntarily provides real-time delay information, for example on a website. For the non-announcer (N), customers are only aware of the periodically updated long-term average delay information. Customers make patronage decisions based on the available delay information. We investigate how A, as the first-mover in announcing real-time information, influences market shares and customer delays. We find that when A is not the higher-capacity (i.e. faster) service provider, real-time announcements improve her performance with respect to increasing market share and decreasing customers’ delays. However when A is the higher-capacity service provider, the effects are mixed, possibly resulting in A losing market share, or her customers experiencing longer delays. We also investigate the effects of A’s announcements on social welfare and N’s performance, and find that delay announcements improve social welfare when A is not the higher-capacity service provider.

Key words: service systems; real-time delay announcements; admission control; quasi-birth-and-death processes

1. Introduction

Some service providers announce information about anticipated delay as a means to influence customers’ patronage decisions and manage congestion. For example, call centers use standard software to announce callers’ expected delay when they dial, sometimes updating callers while they are waiting, to influence their decisions on whether to continue to wait or call back at a later time. Advances in internet-based technology, mobile apps, and managerial practice facilitate conveying delay information to customers even before their physical interaction with a service provider, possibly to help them make informed decisions about which service provider to patronize. In this paper we are interested in the effects of such applications.

In healthcare, publicly available emergency room (ER) wait times can help patients present to a less crowded ER. For example, Florida Hospital publishes real-time expected delays of its 22 ERs on a smart-phone application (Florida Hospital 2016), and the paid ERtexting service allows hospitals to text their expected delay to a central server, which broadcasts the information to
the community (Sadick 2012). Restaurants also take advantage of internet-based applications to manage customer flow. *NoWait*, for example, is a queue management application that disseminates restaurants’ expected time-to-seat, enabling diners to choose one of several restaurants based on their anticipated delays (Perez 2014). Crowdsourcing applications also enable delay information dissemination by allowing previous customers to inform others about their delay experience. For example, the smart-phone application *Waze* crowdsources traffic information (Needleman 2009), *Azneh* and *Qalandiya* crowdsource congestion at various Israeli checkpoints (Hazboun 2015), and *Ride Hopper* crowdsources ride wait times in amusement parks (Blecherman 2012).

When a service provider functions *in isolation*, the Operations Management literature has documented advantages of announcing delay information for both the service provider and the customers (e.g., Whitt 1999). However, the network effects of such initiatives are unknown in a competitive business environment in which customers are willing to receive service from one of multiple service providers, choosing one over the others based on an anticipated shorter wait. In such a setting the benefits of delay information will depend on which service providers announce what type of information. For example, some service providers might announce real-time delay information, whereas others might only provide long-term historical average delays. This is the case, for example, for ERs who are not required to announce real-time information (though some do), but are required to publicize their one-year average wait times in the Center for Medicare and Medicaid Services database (Groeger et al. 2014).

From the viewpoint of a service provider who is competing for market share and considering the option of announcing her real-time delay information, it is crucial to understand the impact of such a decision on her operations and business environment. In this paper we investigate the first-mover’s potential benefit from taking this initiative. We answer this central question: In a competitive environment what benefits might accrue to the service provider who first provides real-time delay information to customers, when the only available information about the other service providers are long-term historical delays?

We propose queueing models of two competing service providers with comparable services, who announce different types of delay information. One service provider initiates announcing truthful real-time expected delay information, while the other provides truthful historical average expected delay information updated at regular intervals, e.g., monthly, quarterly, or yearly. Customers patronize the service provider with the shorter expected delay. We investigate how real-time

---

1 We do not consider systematic falsification of delay information. The backlash from systematically lying could result in loss of customer goodwill and even potentially legal action (see O’Donnell (2014) for a related report about the repercussions of delay falsification at a Veterans’ Administration Hospital)
announcements impact the market share and expected delay of the two service providers. We consider these metrics as opposed to profit, for example, to avoid the need to impose functional forms on the profits of the service providers, or assumptions on their tangible and intangible waiting costs (such as queue management effort, physical queueing space, and loss of goodwill due to long delays). Service managers could employ our analyses and insights to inform their decisions based on their own assessments of the relative importance of the metrics.

We find that the benefits that accrue to a service provider who initiates real-time delay announcements depend on how her service capacity compares to her competitor’s service capacity. In particular, when the service provider who initiates announcements is not the higher-capacity (i.e. faster) provider, she improves her performance with respect to increasing market share and decreasing customers’ delays. However, the impact of such an initiative by a higher-capacity service provider is mixed, possibly resulting in market share loss or longer delays for her customers, depending on the market size and the service providers’ capacities and service rates. This result suggests that the higher-capacity service provider needs to be cautious about revealing real-time delay estimates.

This is in contrast with previous findings in single-service provider settings, where providing more precise delay information always favors the service provider. We also find, not surprisingly, that the impact of announcing real-time delay information is more pronounced as more customers make their patronage decisions based on the provided delay information. We also investigate the societal impacts of one provider initiating real-time delay announcements, and find that social welfare always increases when the announcer benefits on market share and customer delay.

2. Literature Review

We classify the related literature into two streams: (1) papers that study delay estimators, and (2) papers that investigate the impact of announced information on customers’ patronage decisions and system performance. Our work is more relevant to the second stream; we review the literature from the first stream as it helps to introduce terminology that we use later in our model.

2.1. Delay Estimators

It is well documented that in a Markovian first-come-first-served (FCFS) setting with accessible queue length information, the commonly used queue length (QL) estimator is the most accurate delay estimator (accuracy being measured as the mean squared deviation of the actual delay from the estimated delay) (Ibrahim and Whitt 2009a). When the Markovian setting is violated or the queue length information is not accessible, Ibrahim and Whitt (2009a) propose three alternative history-based estimators: (1) delay of the last customer to enter service (DLS), (2) delay experienced so far by the customer at the head of the line (HOL), and (3) delay of the customer who has arrived most recently among those who have completed service. They show that errors of the history-based
estimators are negligible in the many-server and heavy-traffic limiting regimes. Ibrahim and Whitt (2009b) adjust QL and HOL estimators to incorporate customers’ abandonment, and Ibrahim and Whitt (2011) propose estimators for time-varying number of servers and time-varying arrivals. Ibrahim et al. (2015) study how customers’ announcement-dependent balking and abandonment impact the accuracy of the DLS estimator, and prove that DLS prediction is asymptotically accurate under the quality-and-efficiency-driven and the efficiency-driven many-server heavy-traffic regimes if, and only if, the fluctuations in the queue length process are small. Jouini et al. (2009) modify QL estimators for call centers with priority service discipline and impatient customers, and Jouini et al. (2015) develop optimal delay estimators based on an Erlang-distribution approximation for priority systems where customers perceive under and over announcements differently. Ang et al. (2015) propose the Q-Lasso method, which combines queueing-based delay estimators with the Lasso method of statistical learning, to estimate ER wait times. They show that Q-Lasso predictions are more accurate for low-acuity patients compared to rolling average methods, fluid model estimators, and regression methods.

2.2. Impact of Delay Information Provision

We further classify this stream into two categories: delay information provision in the single (Sections 2.2.1) and multiple-service provider (2.2.2) settings.

2.2.1. Delay Information Provision in a Single-Service Provider Setting

Whitt (1999) investigates the performance of a call center when the QL delay estimates are announced to arriving callers, who use them to decide either to balk or join. Armony and Maglaras (2004) model a call center in which callers may balk, choose to wait, or ask for a call back. Both papers document that announcing the QL delay estimator improves the average waiting time and the utilization of the system. Armony et al. (2009) make the balking and abandonment functions endogenous by allowing customers to maximize their expected utility from service and waiting, and characterize the equilibrium behavior associated with announcing the DLS estimate in a many-server queue setting.

A few papers document settings in which inappropriate delay announcement strategies might have adverse impacts on customers and service providers. For example, Guo and Zipkin (2007) explore the effects of varying the precision level in delay announcements (no information, the queue length information, and the accurate waiting time information) on the performance of queueing systems. Customers either join the system or balk, depending on their utility function and the provided information. Although they find that in most settings more information improves system performance, the authors identify some extreme conditions under which more information hurts the service provider’s utilization and customers’ wait times. Allon and Bassamboo (2011) study
the impact of providing delay information *with a lag* (after a customer has waited for some time) instead of providing information immediately upon arrival. In a “cheap talk” setting (in which the information exchange is free and does not directly affect payoffs), they find that the delay in announcement may diminish the firm’s credibility.

Several papers investigate how firms might manipulate delay announcement strategies to induce specific behaviors in customers, and how those strategies might have an adverse impact on customers’ experience. For example, Hassin (1986) shows that a revenue-maximizing service provider might choose not to disclose any information about the system state when delay information is a signal of quality, though doing so may be socially suboptimal. Alternatively, a self-interested firm that wishes to induce a desired behavior from strategic customers may intentionally provide vague information, as studied by Allon et al. (2011). Jouini et al. (2011) analyze the effect of announcing different percentiles of the waiting time distribution (announcement coverage) in a model in which customers balk and renege in reaction to delay announcements. They show how a manager may use announcement coverage as a lever to control the trade-off between the informed balking and misinformed reneging decisions of customers. Hu et al. (2015) study a system with two streams of customers, one informed about real-time delay and the other uninformed, and show that some amount of information heterogeneity in the population can lead to strictly more favorable system throughput or social welfare, as compared to a system with information homogeneity.

2.2.2. Delay Information Provision in a Multiple-Service Provider Setting

An example of delay announcement policies in systems with multiple service providers is the phenomenon of “ambulance diversion” in emergency medical services (EMS). Ambulance diversion policies aim to balance EMS load by allowing ERs (the service providers) to request diversion of incoming ambulances to neighboring hospitals during overcrowding periods. Deo and Gurvich (2011) study centralized and decentralized EMS systems with two ERs employing threshold-based diversion policies, where each hospital is “on diversion” if its census is above a threshold. They establish the Pareto optimality of the defensive equilibrium in which both ERs are always on ambulance diversion in the decentralized setting, and show that a simple centralized control policy is approximately socially optimal and is Pareto improving for both ERs. Do and Shunko (2015) present a threshold policy that is Pareto improving over the decentralized policy for a variety of metrics. This policy is chosen such that the expected arrival rate at each hospital remains identical to what it was before implementation of the policy. Enders (2010) investigates the effect of hospital capacities on threshold policies in a decentralized EMS system with two competing hospitals and illustrates that the “never divert” policy for both hospitals is sometimes optimal. He also evaluates the impact of coordination between the two hospital on quality of care, revenue, and ambulance response times.
In a setting where a neighboring ED’s wait time distribution is known, Ramirez-Nafarrate et al. (2014) find optimal diversion policies that minimize average wait beyond a safety time threshold.

In an empirical study, Dong et al. (2015) observe that patients take delay information, if available, into account when choosing among multiple ERs. In addition, their numerical studies show that delay announcements increase coordination in an ER network and thereby improve patients’ wait times. Similarly, He and Down (2009) show via heavy traffic analysis and simulation that delay announcements improve patients’ wait times even when only a small proportion of the patients are flexible (willing to choose an ER based on a shorter expected delay). They also consider the impact of information updating frequency, finding that flexibility can hurt wait times if information is updated infrequently. Another motivating application for He and Down’s (2009) model is call centers with bi-lingual agents.

In summary, the existing body of literature studies the impact of delay information either in single-service provider settings or in multiple-service provider settings in which all service providers disseminate equally rich delay information. This leaves questions about the first-mover’s advantage from providing richer information in a competitive setting unanswered. We investigate these questions using a model of two comparable service providers in which one of them initiates real-time expected delay announcements as a competitive lever. To the best of our knowledge we are the first to explore such a setting.

3. System Setting and Model Formulation
We consider a system of two competing service providers with comparable services. Customers choose the service provider where they expect to encounter a shorter delay before being served. Initially, neither service provider announces real-time delay information, and therefore, when customers seek to minimize delay, equilibrium arrivals to the service providers will be such that the delays at the service providers are equal. With this as the starting point, we investigate the evolution of the system when one service provider, the Announcer $A$, considers announcing her real-time delay information, while the other, the Non-announcer $N$, only publishes her historical average delay at regular updating periods indexed by $t$, for example, at the end of each month or quarter. Provider $A$ announces real-time delay information starting in Period $t = 1$. We assume that the updating periods are sufficiently long to ensure that the system reaches stationarity (or divergence becomes apparent if the system is unstable) in each period. We also ensure that $A$ and $N$ are truthful in their announcements.

In a general Period $t \geq 1$, customers compare the available information, that is, $A$’s real-time QL delay predictions and $N$’s average delay in Period $t - 1$, and choose the lesser of the two, breaking ties in favor of $N$ for simplicity. Customers’ patronage decisions in Period $t$ affect the performance
of the two service providers, and consequently the information provided by \( N \) in Period \( t + 1 \).

We ignore any gap, e.g. travel time, between the moment that customers make their patronage decisions and their arrival at the chosen service provider to avoid extra modeling complications, and for tractability reasons. This is a reasonable assumption for some applications; for example, the NoWait application has a tool that allows customers to get in line before arriving at a restaurant.

We consider a Poisson process with rate \( \Lambda \) for customer demand, and we denote the average arrival rates to \( A \) and \( N \) in Period \( t \) by \( \lambda_{t}^{(A)} \) and \( \lambda_{t}^{(N)} \), respectively. Note that \( \Lambda = \lambda_{t}^{(A)} + \lambda_{t}^{(N)} \) is an exogenous parameter whereas \( \lambda_{t}^{(A)} \) and \( \lambda_{t}^{(N)} \) are determined based on system dynamics. Both \( A \) and \( N \) are single-server queues with exponential service times with respective rates \( \mu^{(A)} \) and \( \mu^{(N)} \) (we treat the multi-server case in Section 7.1). We denote the average queue delays (time to service) and market shares of \( A \) and \( N \) in Period \( t \) by \( D_{t}(i) \) and \( M_{t}(i) \), \( i \in \{A, N\} \), respectively. Assumption 1 ensures that the system is initially stable (in Period \( t = 0 \)):

**Assumption 1.** \( \Lambda < \mu^{(A)} + \mu^{(N)} \).

In Period \( t = 0 \), \( A \) and \( N \) behave as \( M/M/1 \) queues with equilibrium arrival rates that result in \( D_{0}^{(A)} = D_{0}^{(N)} \) (we derive \( \lambda_{0}^{(A)} \) and \( \lambda_{0}^{(N)} \) in Section 4). From Period \( t = 1 \), \( A \) starts announcing real-time QL delay estimates: If the current number of customers in \( A \) is \( n^{(A)} \), the expected delay of an arriving customer is \( d_{n} = n^{(A)}/\mu^{(A)} \). Without loss of generality, we scale time such that \( \mu^{(A)} = 1 \) for the rest of the paper; therefore, \( d_{n} = n^{(A)} \).

In Period \( t \geq 1 \), customers compare \( d_{n} \) with \( D_{t-1}^{(N)} \) (\( N \’s \) average delay in the previous period), and join \( A \) if \( d_{n} < D_{t-1}^{(N)} \) and \( N \) otherwise. This induces a threshold structure to model arrivals to \( A \) and \( N \) in Period \( t \geq 1 \): When \( n^{(A)} < C_{t} = \lceil D_{t-1}^{(N)} \rceil \), where \( \lceil \cdot \rceil \) is the ceiling function, the arrival rates to \( A \) and \( N \) are \( \Lambda \) and \( 0 \), respectively, and when \( n^{(A)} \geq C_{t} \), the arrival rates to \( A \) and \( N \) are \( 0 \) and \( \Lambda \), respectively. Figs. 1a-1b present the resulting Continuous Time Markov Chain (CTMC) models for \( A \) and \( N \) in Period \( t \geq 1 \), where \( n^{(A)} \) and \( n^{(N)} \) denote the current number of customers in \( A \) and \( N \), respectively.

Model \( A \) is a birth-death process with the states representing \( n^{(A)} \). Model \( N \) with state space \( \{ (n^{(A)}, n^{(N)}): n^{(A)} = 0, \ldots, C_{t}; n^{(N)} = 0, 1, \ldots \} \) is two-dimensional as we need to keep track of both \( n^{(N)} \) and \( n^{(A)} \) to determine the arrival rate to \( N \). Model \( N \) has the following feasible transitions from a general state \( (n^{(A)}, n^{(N)}) \):

- Service completion at \( A \) when \( n^{(A)} > 0 \), with rate \( \mu^{(A)} = 1 \), resulting in a transition to state \( (n^{(A)} - 1, n^{(N)}) \),
- Service completion at \( N \) when \( n^{(N)} > 0 \), with rate \( \mu^{(N)} \), resulting in a transition to state \( (n^{(A)}, n^{(N)} - 1) \),
- Arrival to \( A \) when \( n^{(A)} < C_{t} \), with rate \( \Lambda \), resulting in a transition to state \( (n^{(A)} + 1, n^{(N)}) \),
- Arrival to $N$ when $n^{(A)} = C_t$, with rate $\Lambda$, resulting in a transition to state $(C_t, n^{(N)} + 1)$.

In any Period $t \geq 1$, analysis of Model $N$ determines

$$C_t = \left\lceil D^{(N)}_t \right\rceil,$$

which is used to specify Models $A$ and $N$ in Period $t + 1$. Unlike Model $A$, Model $N$ is not tractable in a general setting: Model $N$ is a Quasi-Birth-and-Death (QBD) process with $n^{(N)}$ as the levels (the infinite dimension), $n^{(A)}$ as the phases (the finite dimension), and skip-free transitions between the levels. We present the matrix blocks of Model $N$ in Appendix D. For the special cases of $C_t = 1$ and $C_t = 2$, we derive closed-form expressions for Model $N$’s delay, which we employ in our analytical derivations in Section 4. For more general settings we employ standard matrix analytic methods (Neuts 1981) to analyze Model $N$ in our numerical experiments in Section 5.

4. Analytical Results for Short and Long-Term Effects of Announcements on $A$’s Performance

In this section we first explain how the traffic initially splits between $A$ and $N$ (in Period $t = 0$). In Section 4.1, we analyze the system one period after $A$ starts announcing real-time information (at the end of Period $t = 1$), to investigate the short-term effects of $A$’s real-time delay announcements on her performance and her customers’ delay experience. In Section 4.2, we solve Model $N$ analytically in Period $t = 2$ under some restricted conditions (as Model $N$ is intractable in a general setting) and we show in which situations we can infer the long-term effect of $A$’s real-time announcements from the short-term analysis.

In Period $t = 0$, neither of the service providers announces delay information, and therefore customers split between $A$ and $N$ such that the expected delays at $A$ and $N$ are identical, i.e., $D^{(A)}_0 = D^{(N)}_0$: $\Lambda$ splits into two independent Poisson streams with intensities $\lambda^{(A)}_0$ and $\lambda^{(N)}_0$. Condition (1)
must hold in order for \( A \) and \( N \) to be stable in Period \( t = 0 \) (Assumption 1 and our assumption that customers split to equalize delay ensure that (1) holds):

\[
\lambda_0^{(A)} < \mu^{(A)} \quad \text{and} \quad \lambda_0^{(N)} < \mu^{(N)}.
\]  

(1)

Thus \( A \) and \( N \) function as independent \( M/M/1 \) queues with expected delays,

\[
D_0^{(A)} = \frac{\lambda_0^{(A)}}{1 - \lambda_0^{(A)}} \quad \text{and} \quad D_0^{(N)} = \frac{\lambda_0^{(N)}}{\mu^{(N)} (\mu^{(N)} - \lambda_0^{(N)})}.
\]  

(2)

By setting \( D_0^{(A)} = D_0^{(N)} \), \( \lambda_0^{(A)} + \lambda_0^{(N)} = \Lambda \), and \( M_0^{(A)} = \lambda_0^{(A)}/\Lambda \), we derive \( A \)'s market share in Period \( t = 0 \):

\[
M_0^{(A)} = \begin{cases} 
\frac{1 + \mu^{(N)} - \Lambda (1 - \mu^{(N)}) - \sqrt{(1 + \mu^{(N)} + \Lambda (1 - \mu^{(N)}))^2 - 4\Lambda (1 - \mu^{(N)})}}{2\Lambda (1 - \mu^{(N)})} & \mu^{(N)} \neq 1, \\
\frac{1}{2} & \mu^{(N)} = 1.
\end{cases}
\]  

(3)

We can similarly derive \( M_0^{(N)} \) using \( \lambda_0^{(N)} = \Lambda - \lambda_0^{(A)} \) and \( M_0^{(N)} = \lambda_0^{(A)}/\Lambda \). Using (2)-(3), \( \lambda_0^{(A)} = M_0^{(A)} \Lambda \), and \( \lambda_0^{(N)} = M_0^{(N)} \Lambda \), the expected delays in Period \( t = 1 \) follow,

\[
D_0^{(A)} = D_0^{(N)} = \begin{cases} 
\frac{1 + \mu^{(N)} - \Lambda (1 + \mu^{(N)}) - \sqrt{(1 + \mu^{(N)} - \Lambda (\mu^{(N)} - 1))^2 + 4\Lambda (\mu^{(N)} - 1)}}{2\mu^{(N)} (\Lambda - \mu^{(N)} - 1)} & \mu^{(N)} \neq 1, \\
\frac{\Lambda}{2 - \Lambda} & \mu^{(N)} = 1.
\end{cases}
\]  

(4)

4.1. Short-term Effect of Delay Announcements

In this section we investigate the short-term effects of \( A \)'s real-time delay announcements by analyzing the system in Period \( t = 1 \) - what happens to \( A \)'s performance in the first period of exercising real-time announcements? We obtain \( A \)'s market share and expected delay in Period \( t = 1 \) in (5)-(6) by solving the balance equations of Model \( A \) (Fig. 1a) where \( C_1 = \lfloor D_0^{(N)} \rfloor \) and \( D_0^{(N)} \) is given in (4):

\[
M_i^{(A)} = \frac{\Lambda \sum_{i=0}^{C_1-1} \pi_i}{\Lambda} = \frac{\Lambda^{C_1} - 1}{\Lambda^{C_1+1} - 1},
\]  

(5)

\[
D_i^{(A)} = \frac{C_1 \Lambda^{C_1}}{\Lambda^{C_1} - 1} + \frac{\Lambda}{1 - \Lambda}.
\]  

(6)

We prove the following proposition about \( A \)'s market shares in Periods \( t = 0 \) and \( t = 1 \) in Appendix A.

**Proposition 1.** \( M_1^{(A)} > M_0^{(A)} \) when \( A \) is the lower-capacity service provider \((\mu^{(N)} > \mu^{(A)})\). \( M_1^{(A)} \geq M_0^{(A)} \) when \( A \) and \( N \) have the same capacity \((\mu^{(N)} = \mu^{(A)})\).
Proposition 1 asserts that A’s real-time delay announcements in Period \( t = 1 \) improve her market share when she is not the higher-capacity (faster) service provider (and weakly improve market share when \( A \) and \( N \) are equally fast). We illustrate Proposition 1 in Figs. 2a-2b, where we compare \( A \)’s market shares in Periods \( t = 0 \) and \( t = 1 \) over the entire feasible range of \( \Lambda \) (\( \Lambda \in (0, \mu(A) + \mu(N)) \)). Note that for a given \( \mu(N) \), the magnitude of improvement depends on \( \Lambda \) in a non-monotonic way. As Fig. 2c shows, \( M_1(A) \geq M_0(A) \) does not hold over the entire \( \Lambda \) range when \( A \) is the higher-capacity service provider.

![Graphs showing market share changes](image)

(a) \( \mu(N) = 2 \) (\( A \) is twice as slow)  
(b) \( \mu(N) = 1 \) (equal service rates)  
(c) \( \mu(N) = 0.5 \) (\( A \) is twice as fast)

**Figure 2**  
Short-Term effects of real-time announcements on \( A \)’s market share

Summarizing Fig. 2, when \( A \) is the lower-capacity service provider (\( \mu(N) > \mu(A) \)), \( M_1(N) > M_0(A) \) and \( M_1(A) > M_0(A) \) (these inequalities are weak when \( A \) and \( N \) have the same capacity). In other words, delay announcements help \( A \) recoup some market share lost to \( N \) in Period \( t = 0 \), because \( A \) is less idle in Period \( t = 1 \). The idle probability of \( A \) is positively influenced by two factors: the rate at which state zero is visited (\( \eta \)) and the expected time spent in state zero per visit (\( T \)). The relative magnitude of \( \eta \) and \( T \) determine \( A \)’s market share. Time \( T \) in Period \( t = 1 \) (\( 1/\Lambda \)) is shorter than \( T \) in Period \( t = 0 \) (\( 1/\lambda_0(A) \)). When \( \mu(N) \geq \mu(A) \), we have \( \Lambda > 2\lambda_0(A) \), resulting in \( T \) being shortened to less than half its value in Period \( t = 0 \). This dominates the effect of \( \eta \) for all values of \( \Lambda \), leading to a smaller idle probability, and hence a larger market share for \( A \).

In contrast, when \( A \) is the higher-capacity service provider (\( \mu(N) < \mu(A) \)), the short-term effect of delay announcements on \( A \)’s market share is mixed, and depends on the service rates \( \mu(N) \) and \( \mu(A) \) and the system load \( \Lambda \). For example in Fig. 2c, \( A \) improves her market share from \( M_0(A) \approx 0.8 \) to \( M_1(A) \approx 0.9 \) when \( \Lambda = 0.1 \). On the other hand, she loses market share from \( M_0(A) \approx 0.76 \) to \( M_1(A) \approx 0.66 \) when \( \Lambda = 0.5 \). In this specific example, \( T \) decreases from 2.62 to 2, but \( \eta \) increases from 0.62 to 1, wiping out any benefit.

The discontinuities in the plots for Period \( t = 1 \) are because of unit changes in the ceiling function used in \( C_1 = \lceil D_0(N) \rceil \). Different ranges of \( \Lambda \) correspond to different \( C_1 \) values: For small values of
Λ, \( C_1 = 1 \) over a range. As we increase \( \Lambda \) beyond that range, \( C_1 \) jumps to 2 and remains at 2 for a range, then jumps to 3, and so on. As \( \Lambda \) increases in its range for a given \( C_1 \), \( M_1^{(A)} \) decreases continuously because \( A \) turns away a larger fraction of customers. When \( \Lambda \) increases more, such that \( C_1 \) increases by one unit, the market share increases discontinuously as \( A \) attracts more customers.

It is worth noting in Figs. 2a-2c that

\[
M_1^{(A)} \rightarrow \begin{cases} 
\mu^{(A)} = 1 & \text{when } \Lambda \rightarrow 0, \\
M_0^{(A)} = \frac{\mu^{(A)}}{\mu^{(A)} + \mu^{(N)}} & \text{when } \Lambda \rightarrow \mu^{(A)} + \mu^{(N)}. 
\end{cases}
\]

When \( \Lambda \rightarrow 0 \), \( C_1 = 1 \) and since traffic is low, \( A \) is rarely in state \( C_1 = 1 \), attracting almost all customers. At the other extreme, when \( \Lambda \rightarrow \mu^{(A)} + \mu^{(N)} \), \( M_1^{(A)} \rightarrow M_0^{(A)} \) - in this case, \( A \) and \( N \) serve at capacity, and their expected delays approach infinity in both periods \( (\lambda_0^{(A)} = \lambda_1^{(A)} \rightarrow \mu^{(A)}) \).

Now we turn our attention to the effect of real-time delay announcements on customers’ average delay at \( A \). We prove Proposition 2 in Appendix B.

**Proposition 2.** \( D_1^{(A)} < D_0^{(A)}, \forall \mu^{(N)} \text{ and } \mu^{(A)}. \)

Proposition 2 asserts that the expected delay at \( A \) improves in Period \( t = 1 \), regardless of the service rate settings. This is simply because announcements enable \( A \) to balance the low-load and high-load epochs by attracting customers when the load is low and diverting them when the load is high.

To summarize, real-time delay announcements improve \( A \)’s market share and delay in the short-term when she is not the higher-capacity service provider. When \( A \) is the higher-capacity service provider, announcements guarantee short-term delay improvement at \( A \), but possibly at the cost of shrinking her market share. This is because \( A \), as the higher-capacity service provider, claims the majority of the market share in Period \( t = 0 \) when customers only access the expected delay information. By announcing real-time congestion information in Period \( t = 1 \), she may end up turning away some of those customers, specifically at times that \( A \) highly congested.

### 4.2. Extending Short-Term Analysis to the Long Term

In Section 4.1 we studied the short-term effects of announcements on \( A \)’s performance. In order to analyze the long-term effects we need to understand how \( C_t \) evolves over time, since Period \( t \) is related to Period \( t - 1 \) through \( C_t = \left[ D_{t-1}^{(N)} \right] \). As a first step, we investigate whether the analytical results of Propositions 1-2 are extendable to the long term by analyzing Period \( t = 2 \), which requires analytical characterization of Model \( N \) (Fig. 1b) in Period \( t = 1 \) to determine \( C_2 \).

To that end we characterize two of the possible patterns in the evolution of \( C_t \): (1) unstable oscillation and (2) convergence. If \( A \)’s real-time announcements in Period \( t = 1 \) cause \( N \) to be *unstable* in Period \( t = 1 \), then thereafter \( A \) will receive all of the traffic and \( N \) will receive none in
even numbered periods, and \( N \) will receive all of the traffic and \( A \) none in odd numbered periods. If \( C_t \) converges in Period \( t = 2 \), \( C_2 = C_1 \), and the same \( C_t \) (and hence system performance) persists \( \forall t > 2 \). If these patterns are established in Period \( t = 2 \), we will be able to predict the long-term evolution of \( C_t \). We now explore the conditions under which these patterns arise.

**Unstable oscillation.** This pattern occurs when \( N \) is unstable in Period \( t = 1 \), resulting in \( D_1^{(N)} = \infty \). In this case, \( A \) receives all the traffic in Period \( t = 2 \), that is, \( \lambda_2^{(A)} = \Lambda \) and \( \lambda_2^{(N)} = 0 \). Since \( C_3 = D_2^{(N)} = 0 \), \( N \) receives all traffic in Period \( t = 3 \), assuming that ties go to \( N \). So, \( A \) and \( N \) attract all the traffic in alternating periods. Although this situation is unlikely in practice (as customers’ abandonment prevents instability, for example), it is theoretically possible to occur in our model, and is indicative of a system that performs poorly. Proposition 3 states the condition under which unstable oscillation begins in period \( t = 1 \). The proof of the proposition is presented in Appendix D.

**Proposition 3.** \( N \) is unstable in Period \( t = 1 \) if and only if \( \lambda_1^{(N)} \geq \mu^{(N)} \), where \( \lambda_1^{(N)} = \Lambda - \lambda_1^{(A)} \) and

\[
\lambda_1^{(A)} = \Lambda \sum_{i=0}^{C_1-1} \pi_i = \frac{\Lambda (\Lambda C_1 - 1)}{\Lambda C_1 + 1 - 1}.
\]

Instability can arise even when the system as a whole is stable in Period \( t = 0 \). In Section 4.1 we demonstrated that \( A \) could lose market share in Period \( t = 1 \) when \( \mu^{(N)} < \mu^{(A)} \) (Fig. 2c, for example); when \( A \) loses too much market share, \( N \) can become unstable. The following corollary provides a sufficient condition for \( N \) to be stable.

**Corollary 1.** When \( \mu^{(N)} \geq \mu^{(A)} \) (\( A \) is not the higher-capacity service provider), \( N \) is guaranteed to be stable in Period \( t = 1 \).

Based on the stability condition (1) for Period \( t = 0 \), \( \lambda_0^{(N)} < \mu^{(N)} \). In addition, \( \lambda_1^{(N)} \leq \lambda_0^{(N)} \) when \( \mu^{(N)} \geq \mu^{(A)} \) (Proposition 1), and therefore \( \lambda_1^{(N)} < \mu^{(N)} \), implying that \( N \) is stable in Period \( t = 1 \) when \( \mu^{(N)} \geq \mu^{(A)} \).

**Convergence.** When \( C_2 = C_1 \), the systems in Periods \( t = 1 \) and \( t = 2 \) are identical, and therefore, \( C_t = C_2 = C_1, \forall t > 2 \), implying that \( C_t \) converges. Analytical characterization of the conditions that give rise to this pattern would inform us of the conditions under which Propositions 1-2 persist in the long term. Although this is not practical for a general \( C_1 \), as \( N \)’s QBD process (Fig. 1b) with an arbitrarily high number of phases is analytically intractable, we are able to solve Model \( N \) analytically and derive closed-form expressions for \( C_2 \) when \( C_1 = 1 \) or 2. This enables us to provide conditions that result in \( C_2 = C_1 \) when \( C_1 = 1 \) or 2.

We present the details of the solution approach in Appendix E. To summarize it, we first solve a simultaneous non-linear system of equations to obtain the stationary probability of state \((C, 0)\),
which we use in a series sum to derive the expected queue length at provider $N$ in Period $t = 1$. Next, we employ Little’s Law to obtain $C_2 = \left[D_1^{(N)}\right]$. Our procedure remains valid for $C_1 > 2$, but it results in higher-order simultaneous non-linear systems of equations, which cannot be solved in closed form. When $C_1 = 1$, we derive

$$C_2 = \left[\frac{\mu^{(N)} + 2\Lambda + 1 - \sqrt{\left(\mu^{(N)} + 1\right)^2 + 4\Lambda}}{\mu^{(N)} - 2\Lambda - 1 + \sqrt{\left(\mu^{(N)} + 1\right)^2 + 4\Lambda}}\right].$$

(8)

Using (4) and (8) and simplifying inequalities $D_0^{(N)} < 1$ and $D_1^{(N)} < 1$, which represent $C_1 = C_2 = 1$, we obtain the following explicit condition for convergence in Period $t = 2$.

**Proposition 4.** Convergence occurs at $C_1 = C_2 = 1$ in Period $t = 2$ if

$$\Lambda < \frac{1}{2} \min\left\{\frac{\mu^{(N)}\mu^{(N)^2} - 1 + \sqrt{(\mu^{(N)} + 1)(\mu^{(N)} + 5)}}{\mu^{(N)} + 1}, \frac{2\mu^{(N)^2} + \mu^{(N)} + 1}{\mu^{(N)} + 1}\right\}.$$

Deriving a similar convergence condition is also possible for $C_1 = C_2 = 2$, but we do not present it here because it is unwieldy. Figs. 3a-3b graphically illustrate the convergence conditions in Period $t = 2$ for $C_1 = C_2 = 1$ and $C_1 = C_2 = 2$, respectively. The regions with thick black boundaries specify the parameter settings for which $C_1 = 1$ (Fig. 3a) and $C_1 = 2$ (Fig. 3b). The shaded regions specify the parameter settings for which the convergence conditions in Period $t = 2$ hold, and therefore, our short-term effect findings in Propositions 1-2 sustain in the long term.

![Figure 3](image-url)  
**Figure 3** Parameter settings that result in convergence in Period $t = 2$

To further understand the shaded regions of Figs. 3a-3b, we plot $C_t$, $D_t^{(A)}$, and $M_t^{(A)}$ as they evolve over six periods for two experiments in Figs. 4a-4b, both demonstrating convergence in
Period \( t = 2 \). The parameters for the Fig. 4a experiment fall in the shaded region of Fig. 3a, meaning \( C_t = 1 \) for all periods, while the parameters for the Fig. 4b experiment fall in the shaded region of Fig. 3b, meaning \( C_t = 2 \) for all periods.

Therefore, A’s long-term delay improvement in the Figs. 4a-4b experiments could be predicted based on Proposition 2 and the shaded regions of Figs. 3a-3b. Likewise, A’s long-term market share improvement in the Fig. 4b experiment could be predicted based on Proposition 1 and the shaded region of Fig. 3b, since A is the lower-capacity (slower) service provider. But, we cannot predict the change in A’s market share for the Fig. 4a experiment, because \( \mu(N) < \mu(A) \) in that experiment. However, the shaded region in Fig. 3a shows that whatever the short-term effect, it will sustain in the long term.

5. Numerical Experiments for Long-Term Effects

In Section 4.2 we analytically characterized the evolution of \( C_t \) and the system in the long term, under restricted parameter settings. In this section we perform extensive numerical experiments to investigate the long-term effects under more general conditions. Specifically, we explore the questions of what sort of stable patterns for the evolution of \( C_t \) can arise (in Section 5.1), and how A’s real-time announcements affect her market share and delay in the long term (in Section 5.2).

In our experiments we vary \( \mu(N) \) over the set \( \{1/6, \ldots, 1/2, 1, 2, \ldots, 6\} \). For each \( \mu(N) \), we choose 200 equally spaced values of \( \Lambda \) in the range \((0, 1 + \mu(N))\), resulting in a total of 2200 experiments. Note that \( \Lambda < \mu(A) + \mu(N) = 1 + \mu(N) \) ensures stability of the system in Period \( t = 0 \) (Assumption 1). We focus below on cases where \( N \) is stable in all periods. When \( \mu(N) \geq \mu(A) \), we showed that \( N \) is stable in Period \( t = 1 \) in Corollary 1. All our experiments with \( \mu(N) \geq \mu(A) \) suggest that this result persists for all future periods. For some values of \( \mu(N) < \mu(A) \), \( N \) can become unstable for some values of \( \Lambda \).
5.1. Stable Configurations

We observe two general stable patterns: (1) convergence and (2) stable oscillation. Convergence arises when \( \exists t > 1 : C_t = C_{t-1} \), and therefore, \( C_t = C_{t'}, \forall t' > t \). We analytically characterized some parameter settings that result in convergence in Period \( t = 2 \) (Figs. 3a-3b). Figs. 5a-5b present two representative experiments with different convergence patterns. In the Fig. 5a experiment, \( C_1 = 2 \). Based on Fig. 3b and the corresponding \( \mu^{(N)} \) and \( \Lambda \), we can predict anything else about the evolution of the system. Fig. 5a shows convergence is established in Period \( t = 5 \). In the Fig. 5b experiment, \( C_1 = 12 \), for which we do not have any analytical results about the long-term pattern.

Figure 5  Convergence pattern

Stable oscillation occurs when \( \exists t > 2 : C_t = C_{t-2} \neq C_{t-1} \), resulting in \( C_t = C_{t'}, \forall t' > t \) (in all our experiments, \( C_t \) oscillated between exactly two levels, not necessarily adjacent). Fig. 6 plots the evolution of \( C_t, D_t^{(A)} \), and \( M_t^{(A)} \) for an experiment with the oscillation pattern being established in Period \( t = 4 \).

5.2. Long-Term Effects of Real-Time Announcements on A’s Market Share and Delay

In this section we discuss in more detail the impacts of announcements on A’s long-term average market share, \( \overline{M}^{(A)} \), and delay, \( \overline{D}^{(A)} \). If convergence is established in Period \( t \), we set \( \overline{M}^{(A)} = M_t^{(A)} \) and \( \overline{D}^{(A)} = D_t^{(A)} \). If stable oscillation is established in Period \( t \), we compute \( \overline{M}^{(A)} (\overline{D}^{(A)}) \) as the average of \( M_t^{(A)} \) and \( M_{t+1}^{(A)} \) (\( D_t^{(A)} \) and \( D_{t+1}^{(A)} \)). For all experiments, we measure \( \Delta M^{(A)} = (\overline{M}^{(A)} - M_0^{(A)})/M_0^{(A)} \) and \( \Delta D^{(A)} = (D_0^{(A)} - \overline{D}^{(A)})/D_0^{(A)} \), the long-term % changes in A’s market share and in customers’ delay at A after initiating real-time announcements (in both measures, a positive value indicates long-term benefits to A).
For each $\mu(N)$ in our experiments, we compute the average of $\Delta M(A)$ and average of $\Delta D(A)$ over the corresponding 200 values of $\Lambda$. We denote the averages by $\overline{\Delta M(A)}$ and $\overline{\Delta D(A)}$ and we plot them in Figs. 7a-7b, respectively (we show the plots up to $\mu(N) = 2$ to keep the figures clear. Our findings and the following explanations hold for $\mu(N) > 2$ as well). The plots show that $\overline{\Delta M(A)}$ and $\overline{\Delta D(A)}$ improve as the ratio $\mu(N)/\mu(A)$ increases, and that $A$ can expect to enjoy long-term benefits in both market share and delay, on average, by announcing her real-time delay information when $\mu(N)/\mu(A) \geq 0.25$, or, if you prefer, when $A$ is no more than four times faster than $N$. When $\mu(N)/\mu(A) < 0.25$, $A$ has to trade off a slight long-term market share gain for a potentially significant delay degradation.

**Figure 6  Oscillation pattern; Parameters: $\mu^{(A)} = 1, \mu^{(N)} = 0.5, \Lambda = 0.87$**

Plots 7a-7b show average long-term changes. One remaining question is how representative these averages are: Are there $\mu^{(N)}$ values for which $A$ can expect performance improvement under any total load $\Lambda$? The answer is “yes.” We find that in all 1200 experiments in which $\mu^{(N)} \geq \mu^{(A)}$ ($A$ is not the higher-capacity service provider), $\Delta M(A) > 0$ and $\Delta D(A) > 0$. For example, Figs. 8a-8b plot $\Delta M(A)$ and $\Delta D(A)$ for all feasible $\Lambda$ values when $\mu^{(N)}/\mu^{(A)} = 2$.

![Figure 7  A's average long-term performance change after real-time delay announcements](image-url)
Propositions 1-2 and an understanding of how $C_t$ evolves when $\mu^{(N)} \geq \mu^{(A)}$ can explain Figs. 8a-8b. Based on Propositions 1-2, real-time delay announcements improve $A$’s short-term market share and delay when $\mu^{(N)} \geq \mu^{(A)}$. Whether the short-term benefits sustain in the long term depend on how $C_t$ evolves in the following periods. Based on Proposition 1 and $M_t^{(N)} = 1 - M_t^{(A)}$, $N$’s market share shrinks in Period $t = 1$ when $\mu^{(N)} \geq \mu^{(A)}$ (as in Fig. 2a, for example). We observe in all experiments in which $\mu^{(N)} \geq \mu^{(A)}$ that $N$’s market share shrinkage in Period $t = 1$ either causes her delay to improve (as observed in all experiments where $\mu^{(N)} > \mu^{(A)}$) or limits the possible degradation to her delay ($N$’s delay degrades in some experiments where $\mu^{(N)} = \mu^{(A)}$). This explains our observation that $C_2 \leq C_1$ in experiments with $\mu^{(N)} \geq \mu^{(A)}$ (Fig. 9 plots average short-term changes to $N$’s delay in such experiments). This immediate effect on $N$’s delay in Period $t = 1$ guarantees that $C_t$ will not stabilize at a level above $C_1$ in the next periods: if $C_t$ converges to $C$, it will be such that $C \leq C_1$, and if $C_t$ oscillates between $C$ and $C'$, it will be such that $\min\{C, C'\} \leq C_1$. For the sake of contradiction suppose that $C_t > C_1$ and $C_{t+1} > C_1$ for some $t \geq 2$; $C_t > C_1$ implies $D_t^{(N)} < D_t^{(A)}$ and $C_{t+1} \leq C_2$. Since $C_2 \leq C_1$, it cannot be the case that $C_{t+1} > C_1$. This prevents $C_t$ from stabilizing above $C_1$. As a result, $A$’s proven short-term delay improvement (Proposition 2) carries over to long-term delay improvement when $\mu^{(N)} \geq \mu^{(A)}$. The fact that $C_t$ does not remain at a level above $C_1$ could result in a drop in $A$’s long-term market share relative to Period $t = 1$. However, we find that the drop is never significant enough to make $A$’s market share worse off in the long term than in Period $t = 0$.

The situation is different when $A$ is the higher-capacity service provider; for some $\Lambda$ values $A$ incurs performance degradation (either in market share or delay) in the long term when $\mu^{(N)} < \mu^{(A)}$. For example, Figs. 10a-10b plot $\Delta M^{(A)}$ and $\Delta D^{(A)}$ for all feasible $\Lambda$ values when $\mu^{(N)}/\mu^{(A)} = 0.25$. Possible long-term market share loss echoes the short-term findings presented in Section 4.1, but long-term delay degradation at $A$ has no precedent in the short-term effects (Proposition 2). This delay degradation at $A$ in the long term can be explained as follows: for all 1000 experiments where $\mu^{(N)} < \mu^{(A)}$, we find that $A$’s real-time delay announcements cause $N$’s delay to degrade in Period
Figure 9  Short-term effect of A’s real-time announcements on N’s delay when \( \mu^{(N)} \geq \mu^{(A)} \)

\[ t = 1 \] (Fig. 11 plots average short-term changes to N’s delay in such experiments). This results in \( C_2 \geq C_1 \), which implies that \( C_1 \) never stabilizes at a level below \( C_1 \) (the reasoning closely follows the discussion for the case \( \mu^{(N)} \geq \mu^{(A)} \)), which explains A’s potential long-term delay degradation. The long-term delay degradation is accompanied by a commensurate increase in long-term market share. However, it is worth noting that in none of the experiments with \( \mu^{(N)} < \mu^{(A)} \) was A worse off on both market share and delay.

Figure 10  Long-term change in A’s performance after real-time delay announcements when \( \mu^{(N)} = 0.25 \)

As Figs. 8 and 10 show, \( \Delta M^{(A)} \) and \( \Delta D^{(A)} \) vary in a discontinuous fashion, forming “regions,” each corresponding to a particular long-term behavior of \( C_t \). In each region, \( C_t \) either converges to a specific value or oscillates between two values. We have labeled the first three regions in Fig. 10. For \( \Lambda \) values in region I, \( C_t \) converges to 1. As \( \Lambda \) increases in region I, A diverts more traffic as she spends more time in the “divert” mode (state \( n^{(A)} = 1 \)), and \( C_t \) is fixed at 1. This causes
\( \Delta M^{(A)} \) to decline, and to possibly be negative. In this region, \( \Delta D^{(A)} = 100\% \), because \( D^{(A)}_t = 0 \) when \( C_t = 1, \forall t \). As \( \Lambda \) increases further \( C_t \) starts oscillating between 1 and 2 (in region II). Again, increasing \( \Lambda \) in this region decreases \( \Delta M^{(A)} \) as \( A \) spends more time in the divert mode (state \( n^{(A)} = 2 \)). This decrease in \( \Delta M^{(A)} \) is accompanied by an increase in \( \Delta D^{(A)} \), because \( A \) diverts comparatively more customers to \( N \) as \( \Lambda \) increases in region II, resulting in smaller relative delays of customers who join \( A \). In region III, \( \Lambda \) is high enough to cause \( C_t \) to converge to 2.

In summary, announcing real-time delay information provides benefits to \( A \) on both market share and delay in the long term when she is not the higher-capacity service provider. The benefits increase as the proportion \( \mu^{(N)}/\mu^{(A)} \) increases. However, the long-term effects might be negative on either market share or delay when \( A \) is the higher-capacity service provider, depending on the system load, \( \Lambda \), and the ratio of service capacities (we found no experiment in which \( A \) is worse off with respect to both market share and delay). Thus, a higher-capacity service provider should carefully evaluate and trade off potential changes to market share and delay before initiating real-time announcements.

### 6. Social Welfare Effects

In the preceding sections, we focused on \( A \)’s performance. In this section we investigate the effects that \( A \), as the sole real-time delay announcer, can have on social welfare—measured as the system-level delay. We compute the system-level delay as the weighted average of customers’ delays at \( A \) and \( N \), with weights being the market shares in each period,

\[
W_t = M^{(A)}_t D^{(A)}_t + M^{(N)}_t D^{(N)}_t,
\]

and \( M^{(N)}_t = 1 - M^{(A)}_t \). From Proposition 1 and the numerical experiments of Section 5, we know that \( N \) loses market share in the short and long term when \( \mu^{(N)} \geq \mu^{(A)} \), and might lose or gain market share when \( \mu^{(N)} < \mu^{(A)} \). Our numerical analysis performed in Section 5 also show that \( N \)’s short and long-term delays can likewise be either better or worse than the Period \( t = 0 \) delay. Therefore, no immediate assertions about the system-level delay can be made.

We compare \( W_0 \) and \( W_1 \) to understand the short-term welfare effects. In Period \( t = 0 \), \( W_0 = D^{(A)}_0 = D^{(N)}_0 \) and \( M^{(N)}_0 = 1 - M^{(A)}_0 \). Proposition 2 asserts that \( D^{(A)}_1 < D^{(A)}_0 \). Clearly, if \( D^{(N)}_1 < D^{(N)}_0 \) (\( A \)’s delay announcements also favors \( N \)’s customers), then \( W_1 < W_0 \) (improvement in system-level delay). Deriving analytical conditions for system-level delay improvement requires us to solve Model \( N \). We obtain these conditions for \( C_1 = 1 \) and 2, based on the derivation of \( D^{(N)}_1 \) in Section 4.2. Using Equation (8), taking \( D^{(N)}_1 < D^{(N)}_0 \) results in a sufficient condition for short-term system-level delay improvement when \( C_1 = 1 \).
**Proposition 5.** The following sufficient condition guarantees short-term system-level delay improvement:

\[
\Lambda < \frac{1}{2} \min \left\{ \frac{2\mu^{(N)}+\mu^{(N)}+1}{\mu^{(N)}+1} \right\},
\]

\[
\max \left\{ \frac{\mu^{(N)}^3 - \mu^{(N)}^2 - \mu^{(N)} + 2 + (\mu^{(N)} - 2)\sqrt{\mu^{(N)}^4 + 2\mu^{(N)}^3 - 5\mu^{(N)}^2 + 2\mu^{(N)} + 1}}{(\mu^{(N)} - 1)^2} \right\} .
\] (9)

Note that condition (9) is sufficient, but not necessary as delay degradation at \( N \) can potentially be offset by improvement in \( A \)'s delay depending on their relative market shares. We also derive a similar condition for \( C_1 = 2 \), but again do not present it because it is unwieldy.

Figs. 12a-12b graphically illustrate the short-term system-level delay improvement conditions for \( C_1 = 1 \) and 2. And the regions with thick black boundaries specify the parameter settings for which \( C_1 = 1 \) (Fig. 12a) and \( C_1 = 2 \) (Fig. 12b). The shaded regions specify the parameter settings where the sufficient system-level delay improvement conditions hold. The regions with dashed boundaries specify parameter settings under which \( C_t \) converges in Period \( t = 2 \), leading to the extension of the short-term performance to the long term. Thus the shaded regions contained within the dashed boundaries specify parameter settings for which system-level delay improves in the short and long term.

![Figure 12](image)

**Figure 12** Conditions for system-level improvement

Deriving general analytical conditions for system-level delay improvement is not possible, even for the short term, because Model \( N \) is intractable for larger \( C_1 \). Instead, we compute \( \Delta_1 W = (W_0 - W_1)/W_0 \), the short-term relative change in system-level delay, for the same numerical experiments as in Section 5. Fig. 13a presents the average of the computed \( \Delta_1 W \) values over all feasible \( \Lambda \) values (\( \Sigma \) for each \( \mu^{(N)} \). According to Fig. 13a, short-term social welfare improves on average by 50% for \( \mu^{(N)} \geq \mu^{(A)} = 1 \). But the average short-term improvement falls sharply when \( \mu^{(N)} < \mu^{(A)} \).
and even becomes negative when \( \mu^{(N)} \leq 1/4 \). Extending the experiments for more periods shows that the long-term effects on system-level delay mirror the short-term ones (Fig. 13b).

![Diagram](image)

**Figure 13** Average change in system-level delay

To understand the intuition behind Figs. 13a-13b, we look into the individual experiments. In all experiments in which \( \mu^{(N)} \geq \mu^{(A)} \) (1200 experiments), we observe that \( \Delta_1 W > 0 \) (short term) and \( \Delta W > 0 \) (long term) for all \( \Lambda \) values; that is, \( A \)’s delay improvement offsets \( N \)’s possible delay degradation when \( A \) is not the higher-capacity service provider.

When \( A \) is the higher-capacity service provider (\( \mu^{(N)} < \mu^{(A)} \)), we observe experiments in which short term system-level delay (\( \Delta_1 W \)) or long-term system level delay (\( \Delta W \)) or both degrade: \( A \)’s real-time delay announcements could hurt the system-level delay (Figs. 14a-14b plot short and long-term system-level delays for all experiments when \( \mu^{(N)} = 0.25 \), for example.) The reason for system-level delay degradation for some systems, in the short-term, is \( N \)’s inability to control her arrivals in Period \( t = 1 \) (unlike \( A \)), which may cause her to perform very poorly with respect to delay (Fig. 13a), particularly when her service rate is low. In the long term, system-level delay degradation can occur because of long-term delay degradation at \( A, N \), or both (Fig. 13b).

It is interesting to note that the discontinuities in system-level delay are less stark in the long term than in the short term (compare Figs. 14a and 14b). This suggests that the long-term level(s) at which \( C_t \) stabilizes serve(s) to counteract the discontinuities, by balancing traffic across \( A \) and \( N \).

In conclusion, social welfare improves in both short and long terms when \( A \) is not the higher-capacity service provider. In addition, we showed in earlier sections that whenever \( A \) is not the higher-capacity provider, she sees market share and delay benefits in both the short and long terms. Accordingly, a service provider who does not have a higher capacity would be inclined to make real-time delay announcements, thereby aligning her incentives with social incentives. In contrast, when \( A \) is the higher-capacity service provider, her decision to announce real-time delay information may lead to degradation in social welfare.
7. Extensions

We study two extensions: multi-server service providers in Section 7.1 and loyal and flexible arrivals in Section 7.2.

7.1. Multiple-Server Service Providers

Models A and N were single-server. In this section we extend our model to consider A and N to be multi-server with $k^{(A)}$ and $k^{(N)}$ servers having service rates $\mu^{(A)}$ and $\mu^{(N)}$ for each server. We retain the same mechanisms for customer routing choices and delay updates. In period $t = 0$, A and N behave as $M/M/k^{(\cdot)}$ systems with equilibrium arrival rates $\lambda_{0}^{(A)}$ and $\lambda_{0}^{(N)}$ being determined such that $D_{0}^{(A)} = D_{0}^{(N)}$. From Period $t = 1$, A announces QL delay estimates: when $n^{(A)} = n, d_{n} = (n - k^{(A)} + 1)^{+}/k^{(A)}$ (again, we set $\mu^{(A)} = 1$). A customer joins A if $d_{n} < D_{t-1}^{(N)}$ and joins N otherwise.

We can restate Proposition 1 for the multi-server case as $M_{1}^{(A)} > M_{0}^{(A)}$ when $k^{(A)} \mu^{(A)} < k^{(N)} \mu^{(N)}$, and $M_{1}^{(A)} \geq M_{0}^{(A)}$ when $k^{(A)} \mu^{(A)} = k^{(N)} \mu^{(N)}$, and Proposition 2 as $D_{1}^{(A)} < D_{0}^{(A)}, \forall k^{(A)}, k^{(N)}, \mu^{(A)}, \mu^{(N)}$. We would expect the propositions to carry over to the multi-server case when $\Lambda$ is relatively high, because a system with $k$ busy servers who serve at rate $\mu$ behaves similarly to a single-server system with service rate $k \mu$ (Artalejo and Lopez-Herrero 2001). Whether the propositions hold when the system load is low is not obvious.

We are able to prove in Appendix C that Proposition 2 continues to hold for the multi-server case. However, proving the multi-server version of Proposition 1 using the same techniques as in...
Appendix A appears impossible, as setting $D_0^{(A)} = D_0^{(N)}$, based on $M/M/s$ formulas and under general parameters, results in an analytically unsolvable equation for $\lambda_0^{(A)}$ and $\lambda_0^{(N)}$ (equilibrium arrival rates in Period $t = 0$). Therefore, we investigate this claim numerically, extending Section 5 set of experiments by varying $k^{(A)}$ and $k^{(N)} \in \{2, 3, 9, 21\}$ (35,200 experiments in total). This allows us to also explore the long-term effects in the multi-server case.

In the short-term, we observe that $M_1^{(A)} > M_0^{(A)}$ in all experiments in which $A$ is not the higher-capacity service provider (18,200 experiments), supporting Proposition 1’s extension to the multi-server case. We also find that the short-term effect on $A$’s market share is mixed when $A$ is the higher-capacity service provider (as it was in the single-server case; Fig. 2c, for example). In the long-term, we find that $\Delta M^{(A)} \geq 0$ and $\Delta D^{(A)} \geq 0$ for all experiments in which $A$ is not the higher-capacity service provider, consistent with the single-server case findings presented in Fig. 7a. Among experiments where $A$ is the higher-capacity service provider, we find some experiments in which either $\Delta M^{(A)} < 0$ (market share degradation) or $\Delta D^{(A)} < 0$ (delay degradation), but there is no single experiment where both market share and delay degradation occur. As in the single-server case, $\Delta_1 W > 0$ and $\Delta W > 0$ when $A$ is not the higher-capacity service provider. However, the social welfare may degrade when $A$ is the higher-capacity service provider.

Thus, the single-server assumption appears to be innocuous, and the findings gleaned from the single-server case continue to be valid: service providers who do not have higher capacity benefit on market share and delay if they announce real-time delay information. However, higher-capacity service providers should carefully evaluate the consequences of announcing real-time delay information on their performance, assessing the trade-off between market share and delay.

7.2. Loyal and Flexible Arrivals

So far we have focused on a system with fully-flexible arrivals: all customers are delay sensitive and make patronage decisions based on delay information. In this section we investigate a system with partially-flexible arrivals, wherein a fraction $p \in (0, 1)$ of customers uses the delay information while the rest of the customers are loyal to one of the service providers.

For the system with partially-flexible customers, we assume single-server service providers (Section 7.1 results showed that the effect of announcements were not sensitive to the number of servers). Just as in the fully-flexible model, we assume customers split between $A$ and $N$ in order to equalize delays in Period $t = 0$, resulting in equilibrium arrival rates of $\lambda_0^{(A)}$ (given in (3)) and $\lambda_0^{(N)} = \Lambda - \lambda_0^{(A)}$. We assume that $p\%$ of each arrival stream consists of flexible customers: $\lambda_0^{(i)}$, $i \in \{A, N\}$, consists of a stream of loyal customers with rate $\Lambda^{(i)} = (1-p)\lambda_0^{(i)}$, and a stream of flexible customers with rate $p\lambda_0^{(i)}$. We denote the total arrival rate of flexible customers to the systems by

\[ 2 \]

Our modeling approach can alternatively be thought of as follows: all customers are fully-flexible before Period $t = 0$ begins, and once delays are equalized in Period $t = 0$, a proportion $(1-p)\%$ become loyal to their choice. This approach ensures that there are enough flexible customers to equalize delays for any given $p$. 
The arrival rates to $A$ and $N$ in Period $t \geq 1$ are $\Lambda^{(A)} + \Lambda^{(f)}$ and $\Lambda^{(N)}$ when $n^{(A)} < C_t$, and $\Lambda^{(A)}$ and $\Lambda^{(N)} + \Lambda^{(f)}$ when $n^{(A)} \geq C_t$. Figs. 15a-15b show the CTMC Models $A'$ and $N'$ in Period $t \geq 1$ in the partially-flexible arrivals case. Both models have infinite state space as loyal customers continue arriving to their service providers regardless of the announced delay information.

We first investigate the sensitivity of the effects of $A$’s real-time announcements on her short-term performance with respect to $p$, and whether Propositions 1-2 hold when $p < 1$. We can obtain $C_1$, which defines Models $A'$ and $N'$ in Period $t = 1$, and derive $D_1^{(A)}, M_1^{(A)}$, and $M_1^{(N)}$ in terms of $C_1, \mu^{(N)}$, and $p$ by analyzing Period $t = 0$. However, the resulting expressions are unwieldy, and they do not allow us to prove analogues to Propositions 1-2 for the partially-flexible arrivals case. Instead, we carry out numerical analysis by extending Section 5 experiments to include 200 equally spaced $p$ values.

When $A$ is not the higher-capacity service provider, $\mu^{(N)} \geq \mu^{(A)}$ (240,000 experiments), we observe that $M_1^{(A)} > M_0^{(A)}$ (in line with Proposition 1), for all values of $p$. However, the benefit of announcements is attenuated as $p$ decreases (see Figs. 16a-16b, for example). When $A$ is the higher-capacity service provider, $\mu^{(N)} < \mu^{(A)}$ (200,000 experiments), again she may gain or lose market share in period $t = 1$ compared to Period $t = 0$ (in line with the fully-flexible arrivals case), as exemplified in Fig. 16c. Note that for some values of $\Lambda$ for which $A$ loses market share in the
short term, partial flexibility can act in A’s favor by mitigating A’s disadvantage (for example, compare $M'_1^{(A)}$ when $p = 0.4$ and when $p = 1$ for $\Lambda = 0.5$ in Fig. 16c).

From Proposition 2, we know that $D_1^{(A)} < D_0^{(A)}$ when $p = 1$. Our experiments show that the proposition continues to hold in the partially-flexible arrivals case ($p < 1$) only when $A$ is the higher-capacity service provider ($\mu^{(N)} < \mu^{(A)}$). As Figs. 16a-16b show $A$ is able to recoup part of her Period $t = 0$ market share loss in Period $t = 1$ when $\mu^{(N)} \geq \mu^{(A)}$. With fully-flexible arrivals, the additional market share in Period $t = 1$ does not increase $A$’s expected delay because she had the ability to turn away the customers who would experience a longer delay than Period $t = 0$ (as $A$ receives no arrivals in state $C_1$). But $A$ loses some of her admission control power when $p < 1$ as she cannot turn away loyal customers; this can result in $D_1^{(A)} > D_0^{(A)}$, especially when the system is more crowded (high $\Lambda$). The problem is exacerbated when values for $\mu^{(N)} > \mu^{(A)}, \Lambda,$ and $p$ are such that the probability of Model $A'$ residing in states $C_1$ and above is larger.

To illustrate this phenomenon, we show in the gray regions in Figs. 17a-17c parameter settings for which real-time delay announcements make $A$ worse off with respect to delay in the short term for three examples with $\mu^{(N)}/\mu^{(A)} = 2, \mu^{(N)}/\mu^{(A)} = 4,$ and $\mu^{(N)}/\mu^{(A)} = 6$. Intuitively, the undesirable short-term effect on $A$’s delay for each $\mu^{(N)}$ arises when $p$ is sufficiently low (fewer flexible customers and more loyal customers) and $\Lambda$ is sufficiently large. Based on our experiments, short-term delay degradation never occurs in the partially-flexible arrivals case under either of the following two conditions for parameter settings: (1) $\Lambda < (\mu^{(N)} - \mu^{(A)})/\mu^{(A)}$, i.e., traffic is quite light, or (2) $p > \mu^{(N)}/(\mu^{(N)} + \mu^{(A)}), i.e., customers are quite flexible.

To investigate the long-term effects in the partially-flexible arrivals case, we ran the experiments for more periods (until the convergence or oscillation pattern is observed) for $p \in \{0.2, 0.4, 0.6, 0.8\}$. Since Model $N'$ is infinite in both dimensions, we need to truncate one of the dimensions to be able to solve it in each period by matrix-analytic algorithms. But, the summation of the state probabilities in each row (phase) of Model $N'$ must match the stationary probability of the corresponding
We find that the long-term effects mimic the short-term effects in the partially-flexible arrivals case: When $A$ is not the higher-capacity service provider, real-time delay announcements increase $A$’s market share at the risk of a potential delay degradation. On the other hand, when $A$ is the higher-capacity service provider, real-time delay announcements always improve her delay, but at the cost of a potential loss of market share. In line with the fully-flexible case, we do not find any experiment where $A$ is worse off on both market share and delay in the partially-flexible arrivals case. In such a setting a higher-capacity service provider $A$ could try to alleviate her market share or delay degradation by making an effort to increase $p$ (making the delay information more accessible and educating the customers to use it).

We find that $A$’s real-time delay announcements improve the short and long-term system-level delay when $A$ is not the higher-capacity service provider, but may degrade short and long-term system-level delay when $A$ is the higher-capacity service provider. He and Down (2009) showed that a higher $p$ improves customers’ overall delay experience in a system with two service providers who both announce real-time delay information. Our experiments suggest a different conclusion when service providers do not disseminate equally rich delay information: a higher $p$ does not necessarily improve the system-level delay, even if the system-level delay at the current level of $p$ is shorter than the Period $t=0$ system-level delay. We find examples of this for values of $\mu^{(N)}$ larger than, smaller than, and equal to $\mu^{(A)}$. For example, Fig. 18 plots the long-term system-level delay ($W$)
in a system where $\mu^{(N)} = \mu^{(A)}$. As the plot shows the effect of $p$ on $W$ is non-monotone in this system: Increasing $p$ from 0.2 to 0.4 decrease $W$, but increasing $p$ further increases $W$.

![Figure 18](image_url)  

**Figure 18** Effect of $p$ on long-term system-level delay when $\mu^{(N)} = \mu^{(A)} = 1$ and $\Lambda = 1.53$

8. Concluding Remarks

The existing body of literature on the effects of delay announcements studies the performance of service systems either in setting with a single service provider, or in settings with multiple service providers who disseminate equally rich delay information to customers. In this paper we modeled and analyzed a setting in which one of two service providers (the announcer $A$) decides to announce real-time delay information, while only periodically updated historical average delay information is available for the non-announcer $N$. Using queueing models, we investigated $A$’s possible first-mover advantage in announcing real-time delay information in terms of improving her market share and the mean delay of her customers.

In the setting with single-server service providers, we proved that $A$ sees benefits in both market-share and delay in the short term (immediately following the initiation of real-time announcements) when she is not the higher-capacity service provider. We showed through numerical experiments that these benefits carry over in the long term, resulting in the system-level delay to improve as well. When $A$ is the higher-capacity service provider, $A$ may lose or gain market share in the short and long terms. In this case, $A$ again improves her performance with respect to delay in the short term, but the improvements may not sustain in the long term. However, $A$ is never worse off with respect to both market share and delay in the long term. The social welfare may either improve or degrade in the short and long terms when $A$ is the higher-capacity service provider.

Our findings are robust to the setting in which service providers are multiple-server, and to setting in which only a proportion of customers use delay information to make patronage decisions, as long as a sufficient number of customers are delay sensitive. When customers are partially-flexible and service providers do not disseminate equally rich delay information, our experiments suggest
a different conclusion from the study by He and Down (2009) in which both service providers announce real-time information: a higher proportion of flexible customers does not necessarily improve system-level delay and may induce a longer delay at the higher-capacity service provider $A$ when service providers do not announce equally rich delay information.

Our work uncovers an effect of making real-time delay announcements that had not been noted previously: making real-time delay announcements is necessarily a good idea only for service providers who do not have relatively high capacity compared to their competitor. However, higher-capacity service providers must exercise caution before providing real-time delay announcements, as this could cause them to either lose market share or cause their customers to experience longer delays. Such service providers must carefully weigh the relative importance of market share and customer delay to their business.

We have not considered $N$’s possible strategic response to $A$’s real-time delay announcements in this paper: It is conceivable that $N$ would initiate real-time announcements in response to $A$’s decision, altering the system dynamics. While this setting is indeed interesting, and has been partially studied in He and Down (2009), our paper seeks to establish the first-mover’s advantage in announcing real-time delay information in the interim between $A$ initiating announcements and $N$ responding.

Appendix A: Proof for Proposition 1

To prove Proposition 1, we show that when $\mu^{(A)} = 1 < \mu^{(N)}$, $M_1^{(A)} > M_0^{(A)}$. The proof that $M_1^{(A)} \geq M_0^{(A)}$ when $\mu^{(A)} = \mu^{(N)}$ closely follows this proof. We aim to show that $\lambda_1^{(A)} > \lambda_0^{(A)} \forall \mu^{(N)} > \mu^{(A)}$. Let $\pi_0^i$ and $\pi_1^i$ denote the probabilities of $A$ being in state $i$ in Periods $t=0$ and $t=1$, respectively. $A$ is an $M/M/1$ queue in Period $t=0$ and an $M/M/1/C_1$ queue in Period $t=1$. From the balance equations and $\mu^{(A)} = 1$:

\begin{align}
\lambda_1^{(A)} &= 1 - \pi_1^0, \\
\lambda_0^{(A)} &= 1 - \pi_0^1.
\end{align}

(10) (11)

To prove $\lambda_1^{(A)} > \lambda_0^{(A)}$, we equivalently prove $\pi_1^0 < \pi_0^0$ (based on (10)-(11)). We accomplish this by contradiction. We divide the feasible $\lambda_0^{(A)}$ values into two regions: $0 < \lambda_0^{(A)} \leq 0.5$ (Case 1) and $0.5 < \lambda_0^{(A)} < 1$ (Case 2). Cases 1-2 cover the entire set of feasible values for $\lambda_0^{(A)}$ (Condition (1)). In our proof, we use the following inequality, which holds because $\mu^{(N)} > \mu^{(A)}$:

\begin{align}
\lambda_0^{(A)} < \frac{\Lambda}{2}.
\end{align}

(12)

Case 1 ($0 < \lambda_0^{(A)} \leq 0.5$): Let’s assume, for the sake of contradiction, that $\pi_1^0 \geq \pi_0^0$, for some $\mu^{(N)} > \mu^{(A)}$. We prove that this assumption leads to the total probabilities of states 0 and 1 in Period $t=1$ being greater than one, $\pi_0^1 + \pi_1^1 > 1$, resulting in contradiction.

\begin{align}
\pi_0^1 + \pi_1^1 &= \pi_0^1 + \pi_0^1 \Lambda \\
&= \pi_0^1 (1 + \Lambda) \geq \pi_0^0 (1 + \Lambda)
\end{align}

, from the balance equations, , from the contradiction assumption,
In this case we prove the contradiction assumption results in $$\pi_0^0 \left(1 + 2\lambda_0^{(A)}\right) \geq (1 - \lambda_0^{(A)}) \left(1 + 2\lambda_0^{(A)}\right)_{LB_1}$$, from (12),

$$= \left(1 - \lambda_0^{(A)}\right) \left(1 + 2\lambda_0^{(A)}\right)_{LB_1}$$, from (11).

We now show that $$LB_1 \geq 1$$, which completes the proof for Case 1. At the two $$\lambda_0^{(A)}$$ extremes in Case 1, $$\lambda_0^{(A)} = 0$$ and $$\lambda_0^{(A)} = 0.5$$, $$LB_1 = 1$$. For $$0 < \lambda_0^{(A)} < 1$$, the first derivative of $$LB_1$$ with respect to $$\lambda_0^{(A)}$$ confirms that $$LB_1 \geq 1$$ because,

$$\frac{dLB_1}{d\lambda_0^{(A)}} = 1 - 4\lambda_0^{(A)} \Rightarrow \left\{ \begin{array}{l} \text{LB}_1 \text{ is increasing in } \lambda_0^{(A)} \text{ when } 0 < \lambda_0^{(A)} < 0.25, \\ \text{LB}_1 \text{ is decreasing in } \lambda_0^{(A)} \text{ when } 0.25 \leq \lambda_0^{(A)} \leq 0.5. \end{array} \right.$$

**Case 2** ($$0.5 < \lambda_0^{(A)} < 1$$): Again, let’s assume, for the sake of contradiction, that $$\pi_0^1 \geq \pi_0^0$$, for some $$\mu^{(N)} > \mu^{(A)}$$.

In this case we prove the contradiction assumption results in $$\pi_0^1 + \pi_1^1 + \cdots + \pi_{C_1}^1 > 1$$:

$$\pi_0^1 + \pi_1^1 + \cdots + \pi_{C_1}^1 >= \pi_0^1 (1 + \Lambda + \cdots + \Lambda^{C_1}) = \pi_0^1 \frac{\Lambda^{C_1+1} - 1}{\Lambda - 1}$$, from the balance equations and the geometric summation,

$$\geq \pi_0^0 \frac{\Lambda^{C_1+1} - 1}{\Lambda - 1}$$, from the contradiction assumption,

$$> \pi_0^0 \frac{(2\lambda_0^{(A)})^{C_1+1} - 1}{2\lambda_0^{(A)} - 1}$$, from (12) and the fact that the geometric sum is increasing in $$\Lambda$$,

$$= \left(1 - \lambda_0^{(A)}\right) \frac{(2\lambda_0^{(A)})^{C_1+1} - 1}{2\lambda_0^{(A)} - 1}_{LB_2}$$, from (11).

We now show that $$LB_2 \geq 1$$, which simplifies to

$$\left(2\lambda_0^{(A)}\right)^{C_1+1} \geq \frac{\lambda_0^{(A)}}{1 - \lambda_0^{(A)}}.$$ (13)

In this case $$2\lambda_0^{(A)} > 1$$. Therefore, we can obtain a lower bound on the left hand side of (13) by finding a lower bound for its exponent $$C_1 + 1$$:

$$C_1 + 1 = \left\lceil \frac{\lambda_0^{(A)}}{1 - \lambda_0^{(A)}} \right\rceil + 1 \geq \frac{\lambda_0^{(A)}}{1 - \lambda_0^{(A)}} + 1 = \frac{1}{1 - \lambda_0^{(A)}}.$$  

Therefore to prove (13), it suffices that we prove

$$\left(2\lambda_0^{(A)}\right)^{C_1+1} \geq \frac{\lambda_0^{(A)}}{1 - \lambda_0^{(A)}} \Leftrightarrow \frac{1}{1 - \lambda_0^{(A)}} \ln \left(2\lambda_0^{(A)}\right) \geq \ln \left(\frac{\lambda_0^{(A)}}{1 - \lambda_0^{(A)}}\right).$$ (14)

When $$\lambda_0^{(A)} \to 0.5$$ (the minimum value for $$\lambda_0^{(A)}$$ in Case 2), $$LHS \to 0$$ and $$RHS \to 0$$. Now to prove (14) we show that $$LHS$$ grows at least as fast as $$RHS$$ as $$\lambda_0^{(A)}$$ increases; that is, we need to show

$$\frac{dLHS}{d\lambda_0^{(A)}} = \frac{1 - \lambda_0^{(A)} + \lambda_0^{(A)} \ln \left(2\lambda_0^{(A)}\right)}{(1 - \lambda_0^{(A)})^2 \lambda_0^{(A)}} \geq \frac{dRHS}{d\lambda_0^{(A)}} = \frac{1}{(1 - \lambda_0^{(A)}) \lambda_0^{(A)}} \Leftrightarrow$$

$$1 - \lambda_0^{(A)} + \lambda_0^{(A)} \ln \left(2\lambda_0^{(A)}\right) \geq 1 - \lambda_0^{(A)} \Leftrightarrow \lambda_0^{(A)} \ln \left(2\lambda_0^{(A)}\right) \geq 0,$$

which holds because $$2\lambda_0^{(A)} > 1$$. This completes the proof for Case 2.
Appendix B: Deriving $D_1^{(A)}$ and the proof for Proposition 2

We first derive $D_1^{(A)}$. Let $\pi_i$ denote the stationary probability of being in state $i$ in Model A (Fig. 1a). Given $\mu^{(A)} = 1$, we use Little’s law to derive the time average delay at $A$ in Period $t = 1$, $D_1^{(A)}$, based on the average queue length, $L_1^{(A)}$, as:

$$D_1^{(A)} = \frac{L_1^{(A)}}{\lambda_1^{(A)}} = \sum_{i=1}^{C_1} \frac{(i-1)\pi_i}{\lambda_1^{(A)}} = \frac{\sum_{i=1}^{C_1} (i-1)\pi_{i-1}A}{\lambda(1-\pi_{C_1})} = \frac{1}{1-\pi_{C_1}} \sum_{i=0}^{C_1-1} i\pi_i,$$  \hspace{1cm} (15)

which eventually simplifies to (6) using $\pi_i = \pi_{C_1}/\Lambda^{C_1-i}$, where $\pi_{C_1} = (\Lambda-1)\Lambda^{C_1}/(\Lambda^{C_1+1} - 1)$ is determined using the balance equations and $\sum_{i=0}^{C_1} \pi_i = 1$.

We now prove Proposition 2. According to (15), $D_1^{(A)}$ is the weighted average of integers $0, 1, \ldots, C_1 - 1$ with weights being $\pi_i/(1 - \pi_{C_1})$, $i = 0, 1, \ldots, C_1 - 1$ (because $\sum_{i=0}^{C_1-1} \pi_i = 1 - \pi_{C_1}$). Therefore,

$$D_1^{(A)} \leq C_1 - 1.$$  \hspace{1cm} (16)

Since $C_1 = \left[ D_0^{(A)} \right]$, we have

$$D_0^{(A)} > C_1 - 1.$$  \hspace{1cm} (17)

Inequalities (16) and (17) prove $D_1^{(A)} < D_0^{(A)}$.

Appendix C: Proof for Proposition 2 for multiple servers

This proof follows similar steps used to prove Proposition 2 in the single-server case. Using Little’s law and setting $k^{(A)} = k$,

$$D_1^{(A)} = \frac{L_1^{(A)}}{\lambda_1^{(A)}} = \frac{\sum_{i=k+1}^{C_1} (i-k)\pi_i}{\lambda_1^{(A)}} = \frac{\sum_{i=k+1}^{C_1} (i-k)\pi_{i-1}A/k}{\lambda(1-\pi_{C_1})} = \frac{1}{1-\pi_{C_1}} \sum_{i=0}^{k-1} \frac{0}{k} \pi_i + \sum_{i=k}^{C_1-1} \frac{i-k+1}{k} \pi_i,$$  \hspace{1cm} (18)

which is the weighted average of $0/k$ (with weights being $\sum_{i=0}^{k-1} \pi_i/(1-\pi_{C_1})$) and $1/k, \ldots, (C_1 - k)/k$ (with weight being $\pi_i/(1 - \pi_{C_1})$, $i = k, \ldots, C_1 - 1$) (because $\sum_{i=0}^{C_1-1} \pi_i = 1 - \pi_{C_1}$). Therefore,

$$D_1^{(A)} \leq \frac{C_1 - k}{k}.$$  \hspace{1cm} (19)

An arriving customer experiences an expected delay of $d_n = (n-k+1)/k$, if the current number of customers in $A$ is $n$. If $d_n < D_0^{(N)} = D_0^{(A)}$ (since ties are broken in favor of $N$), the arriving customer goes to $A$, and otherwise, goes to $N$. The maximum delay that an arriving customer to $A$ might experience in Period $t = 1$ is when the customer finds $C_1 - 1$ waiting customers in $A$ upon arrival (there is no arrival to $A$ when $n = C_1$ in Period $t = 1$). Substituting $n = C_1 - 1$ in $d_n < D_0^{(A)}$, we obtain

$$D_0^{(A)} > \frac{C_1 - k}{k}.$$  \hspace{1cm} (20)

Inequalities (19) and (20) prove $D_1^{(A)} < D_0^{(A)}$ for the multi-server case.
Appendix D: Matrix Blocks for Model $N$ and Proof of Proposition 3

Model $N$ (Fig. 1b) has the following block-tridiagonal transition matrix where all blocks are square matrices of order $C + 1$:

$$Q = \begin{pmatrix} B & A_0 & A_1 & A_2 & \cdots & A_{C-1} & A_C \\ A_2 & A_1 & A_0 & & & & \\ & A_2 & A_1 & A_0 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & \ast & \ast & \\ & & & & 1 & \ast & \ast \\ & & & & & \ast & \ast & \ast \end{pmatrix},$$

where

$$B = \begin{pmatrix} * & \Lambda & & & & & \\ 1 & * & \Lambda & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & 1 & \ast & \ast & & \\ & & & 1 & \ast & \ast & \\ & & & & 1 & \ast & \ast \end{pmatrix}, \quad A_0 = \begin{pmatrix} 0 & & & & & & \\ & \ddots & & & & & \\ & & 0 & & & & \\ & & & \Lambda & & & \\ & & & & 1 & \ast & \ast & \\ & & & & & 1 & \ast & \ast \end{pmatrix}, \quad A_1 = \begin{pmatrix} * & \Lambda & & & & & \\ 1 & * & \Lambda & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & 1 & \ast & \ast & & \\ & & & 1 & \ast & \ast & \\ & & & & 1 & \ast & \ast \end{pmatrix}, \quad A_2 = \begin{pmatrix} \mu^{(N)} & \cdots & \cdots & \cdots & \cdots & \cdots & \mu^{(N)} \end{pmatrix},$$

where “$*$” represents diagonal elements whose values are set such that $Q$ has zero row sums.

We now prove Proposition 3. Model $N$, which is a QBD process, is stable if and only if the following ergodicity condition is satisfied (Latouche and Ramaswami 1999, Theorem 7.2.4):

$$\nu A_0 1 < \nu A_2 1,$$

where the row vector $\nu = (\nu_0, \nu_1, \ldots, \nu_{C+1})$ contains the steady-state probabilities of the Markov chain that corresponds to the generator matrix $A = A_0 + A_1 + A_2$. Matrix $A$ corresponds to the generator matrix of Model $A$ (Fig. 1a). Inequality (21) simplifies to $\nu C_1 \Lambda < \mu^{(N)}$, where $\nu C_1 \Lambda = \lambda^{(N)}_1$, the expected arrival rate to $N$.

Appendix E: Closed-form solutions for $D_1^{(N)}$ when $C_1 = 1$ and 2

Let $\pi_{i,j}$, $i = 0, \ldots, C_1$ and $j \geq 0$, represent the stationary probabilities of Model $N$ in Period $t = 1$. Let $\bar{\pi}_j = (\pi_{0,j}, \ldots, \pi_{C_1,j})$. Model $N$ (Fig. 1b) is a QBD process that is characterized by its square rate matrix $R$ of order $C_1 + 1$, which satisfies $\bar{\pi}_{j+1} = \bar{\pi}_j \times R, \forall j \geq 0$. In an excursion from Level $j$ to Level $j + 1$ initiated by a transition from state $(m,j)$, $m = 0, \ldots, C_1$, element $R_{m+1,n}$ in $R$, $n = 1, \ldots, C_1 + 1$, represents the amount of time the excursion spends in state $(n+1,j+1)$ for every unit of time spent in state $(m,j)$ (Latouche and Ramaswami 1999). Since transitions from Level $j$ to Level $j + 1$ only occur through Phase $C_1$, $R$ contains non-zero entries only in its bottom row $C_1 + 1$. Based on this special structure, $\bar{\pi}_{j+1} = \bar{\pi}_j \times R$ yields,

$$\frac{\pi_{i,j+1}}{\pi_{C_1,j+1}} = \frac{\pi_{i,j+2}}{\pi_{C_1,j+1}}, \quad 0 \leq i \leq C_1, \forall j \geq 0. \quad (22)$$

Let $r_j = \pi_{C_1,j}/(\sum_{i=1}^{C_1} \pi_{i,j})$. From (22), $r_j = r_{j+1} = r, \forall j > 0$. From the flow balance equations, we derive,

$$\lambda^{(N)}_1 = \mu^{(N)} \Pr(N \text{ is busy}) \Rightarrow \lambda^{(N)}_1 = \mu^{(N)} \left(1 - \sum_{i=0}^{C_1} \pi_{i,0}\right) \Rightarrow \sum_{i=0}^{C_1} \pi_{i,0} = 1 - \frac{\lambda^{(N)}_1}{\mu^{(N)}}, \quad (23)$$

where $\lambda^{(N)}_1$ follows $\lambda^{(N)}_1 = \Lambda - \lambda^{(A)}_1$ and equation (7). Taking flow balance equations across levels,

$$\pi_{C_1,j} \Lambda = \mu^{(N)} \sum_{i=0}^{C_1} \pi_{i,j+1} \Rightarrow \sum_{i=0}^{C_1} \pi_{i,j+1} = \frac{\pi_{C_1,j} \Lambda}{\mu^{(N)}}.$$
The expected number of customers in $N$, which we denote by $L_1^{(N)}$, can be expressed as,

$$L_1^{(N)} = \sum_{j=1}^{\infty} j \sum_{i=0}^{C_1} \pi_{i,j} = \sum_{j=1}^{\infty} j \pi_{C_1,j-1} \frac{\Lambda}{\mu^{(N)}} = \sum_{j=1}^{\infty} \left( r_j \sum_{i=0}^{C_1} \pi_{i,j} \right) \frac{\Lambda}{\mu^{(N)}} = \sum_{j=1}^{\infty} \left( r_j \pi_{C_1,j-2} \frac{\Lambda}{\mu^{(N)}} \right) \frac{\Lambda}{\mu^{(N)}} = \sum_{j=1}^{\infty} \sum_{i=0}^{j} \pi_{C_1,0} \left( \frac{\Lambda}{\mu^{(N)}} \right)^j r_j^{j-1}.$$  \hfill (24)

If we can derive the expressions for $\pi_{C_1,0}$ and $r_j$, we can apply Little’s Law and (24) to derive $D_1^{(N)}$.

Let $y$ denote the probability of $A$ being in state $C_1$, which we can easily derive from Model $A$’s balance equations. On the other hand, $y$ can be expressed as a weighted average of $r_j$ values as follows:

$$y = \sum_{j=0}^{\infty} \pi_{C_1,j} = r_0 \sum_{i=0}^{C_1} \pi_{i,0} + \sum_{j=1}^{\infty} r_j \pi_{i,j} = r_0 \left( 1 - \frac{\Lambda_{(N)}}{\mu^{(N)}} \right) + r \frac{\Lambda_{(N)}}{\mu^{(N)}}.$$  \hfill (25)

Therefore, (25) can be used to find $r$ values based on $r_0$. We can find $r_0$ when $C_1 = 1$ and 2 by solving a system of simultaneous non-linear equations governing the stationary probabilities in Levels $j = 0, 1, 2$ of Model $N$. To find probabilities of states in Levels 1-3 of Model $N$, we need a system of $3(C_1 + 1)$ equations. We obtain $2(C_1 + 1)$ equations from flow balance equations for states in Levels 1 and 2, one equation from the sum of Level $j = 0$ state probabilities from (23), and $C_1$ independent equations from (22).

For example, Fig. 19 shows Model $N$ when $C_1 = 1$. The linear equation set (26) lists the flow balance equations for states in Levels 1 and 2 (four equations and six unknowns). We then add the linear equation (27), using (23), and the non-linear equation (28), using (22), which result in an equation set with six equations and unknowns.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure19.png}
\caption{Model $N$ when $C_1 = 1$}
\end{figure}

\begin{align*}
\Lambda \pi_{0,0} &= \mu^{(N)} \pi_{0,1} + \pi_{1,0} \\
(1 + \Lambda) \pi_{1,0} &= \Lambda \pi_{0,0} + \mu^{(N)} \pi_{1,1} \\
(\Lambda + \mu^{(N)}) \pi_{0,1} &= \pi_{1,1} + \mu^{(N)} \pi_{0,2} \\
(1 + \Lambda + \mu^{(N)}) \pi_{1,1} &= \Lambda \pi_{1,0} + \Lambda \pi_{0,1} + \mu^{(N)} \pi_{1,2} \\
\pi_{0,0} + \pi_{1,0} &= 1 - \frac{\Lambda_{(N)}}{\mu^{(N)}} \\
\frac{\pi_{0,1}}{\pi_{1,0}} &= \frac{\pi_{0,2}}{\pi_{1,1}}.
\end{align*}

\hfill (26)

\hfill (27)

\hfill (28)

Solving equations (26) and (27) simultaneously in terms of $\pi_{1,0}$, we obtain the following quadratic equation:

$$\pi_{1,0}^2 \left( \Lambda^2 - (\Lambda + 1) \mu^{(N)} \right) \left( \Lambda (\Lambda + 1) \mu^{(N)} - (\Lambda + 1) \mu^{(N)} \right) + \pi_{1,0} \left( \Lambda^2 - (\Lambda + 1) \mu^{(N)} \right) \times \left( 2(\Lambda + 1) \Lambda^2 \mu^{(N)} - (\Lambda + 1) \Lambda \mu^{(N)} - (\Lambda + 1) \Lambda \mu^{(N)} + (\Lambda + 1) \Lambda \mu^{(N)} \right) + \left( \Lambda^2 - (\Lambda + 1) \mu^{(N)} \right) \left( \Lambda^2 - (\Lambda + 1) \mu^{(N)} \right) = 0,$$

with roots $\Lambda \left( \mu^{(N)} - 2 \Lambda - 1 + \sqrt{4 \Lambda + \mu^{(N)} \left( \mu^{(N)} + 2 \right) + 1} \right) / \left( 2(\Lambda + 1) \mu^{(N)} \right)$. The second root is not admissible because the expression $\mu^{(N)} - 2 \Lambda - 1 - \sqrt{4 \Lambda + \mu^{(N)} \left( \mu^{(N)} + 2 \right) + 1} < 0$ as its value is -2 at $\Lambda = 0$, and it is decreasing in $\Lambda$. Therefore, $\pi_{1,0} = \Lambda \left( \mu^{(N)} - 2 \Lambda - 1 + \sqrt{4 \Lambda + \mu^{(N)} \left( \mu^{(N)} + 2 \right) + 1} \right) / \left( 2(\Lambda + 1) \mu^{(N)} \right)$. We obtain $r_0$ by dividing $\pi_{1,0}$ by $\sum_{i=0}^{\Lambda} \pi_{i,0}$ (given in (23)). We obtain the expression for $r$ by plugging $r_0$ into (25). We then obtain $L_1^{(N)}$ using (24). Finally, we employ Little’s Law to derive (8). The expression for the case of $C_1 = 2$ follows a similar procedure.
References


Do, Hung, Masha Shunko. 2015. Pareto-improving coordination policies in queueing systems with independent service agents. *Available at SSRN 2351965*.


Groeger, Lena, Mike Tigas, Sisi Wei. 2014. ER wait watcher - which emergency room will see you the fastest? URL [https://projects.propublica.org/emergency/](https://projects.propublica.org/emergency/). [accessed on June 26, 2016].


Ibrahim, Rouba, Mor Armony, Achal Bassamboo. 2015. Does the past predict the future? the case of delay announcements in service systems.


Perez, Sarah. 2014. Nowait, the app that lets you join a restaurant wait list from your phone, goes nationwide. URL http://tcrn.ch/1fzSUsc. [accessed on June 26, 2016].
