Title
Models of Bus Queueing at Isolated Bus Stops

Author
Gu, Weihua

Publication Date
2012

Peer reviewed|Thesis/dissertation

Models of Bus Queueing at Isolated Bus Stops

by

Weihua Gu

A dissertation submitted in partial satisfaction of the
requirements for the degree of

Doctor of Philosophy
in

Engineering – Civil and Environmental Engineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Michael J. Cassidy, Co-chair
Dr. Yuwei Li, Co-chair
Professor Carlos F. Daganzo
Professor Rhonda Righter

Spring 2012

Models of Bus Queueing at Isolated Bus Stops

by

Weihua Gu

Abstract

Models of Bus Queueing at Isolated Bus Stops

by

Weihua Gu

Doctor of Philosophy in Engineering – Civil and Environmental Engineering

University of California, Berkeley

Professor Michael J. Cassidy, Co-chair
Dr. Yuwei Li, Co-chair

Long bus queues can form at a busy bus stop where multiple routes converge. These queues and the delays that they impart to buses can often be mitigated by altering the stop's design or the manner in which it is operated. Queueing models are formulated to this end. The models estimate the maximum rates that buses can discharge from a stop that is isolated from other stops and from nearby traffic signals, while maintaining targets for service quality. Model inputs include the stop's number of serial berths, and the processes by which buses arrive and dwell at the stop. The models can be used for a variety of purposes that include: choosing a stop's suitable number of berths or determining when it is useful to alter the rules governing how buses are to dwell in those stops. Application of these models indicates that much of the previous literature on the subject of bus queueing at stops is flawed. And because the analytical solutions derived for the present models describe the special operating features of serial bus berths, the models can be applied to other serial queueing systems common in transport. These include taxi queues, Personal Rapid Transit systems and toll plazas with tandem service booths.

To my parents and my wife, Ruozhu,
for all their love, and everything they have done for me.

Contents

List of Figures

Acknowledgements

The five years I have spent at Berkeley are among the best periods ever in my life, and I appreciate the help and friendship I have received from many people. Without their support and encouragement, I would not be able to make this dissertation possible.

All my success, if any, should first be attributed to my advisor, Michael Cassidy. I have been greatly impressed by his agile mind and superior writing and presentation skills, from which I aspire to learn. As a highly responsible mentor, he has done everything he could to guide my research and help improve my writing and presentations, including this dissertation. Besides, he is very sincere and modest, a true friend in my life. I have learned so much from him, not only in research techniques and skills, but also in scholarliness. My gratitude to him is beyond any words.

I shall deliver the same thanks to my co-advisor, Yuwei Li. He is full of inspiring ideas and has helped me a lot in developing research topics, including my thesis. It has been a joy to work with him. Being one of my best friends, his genuine support also manifests in my daily life at Berkeley.

My acknowledgement also goes to other Berkeley faculty for their guidance and support through my study here. In particular, I want to deliver this acknowledgement to Carlos Daganzo, my committee member and the professor who has deeply shaped my study, research, and ways of thinking. The knowledge, skills, and thoughts that I have learned from him have become my precious wealth. It is my great honor to work with him. I also want to thank Rhonda Righter, the other committee member, and Samer Madanat, for their valuable advice on my research. The same gratitude shall be given to other faculty of the Transportation Engineering program at Berkeley, and my former colleagues at CCIT.

Thanks will also be given to my good friends and classmates at Berkeley. Firstly to Eric Gonzales, who had helped me so much before he left for his new career, especially with my coursework and adaptation to the new environment in my first year. I am grateful to my fellows, Yiguang Xuan and Bo Zou, for their genuine caring and friendship. The list will also include Karthik Sivakumaran, Vikash Gayah, Haoyu Chen, Zhen Sun, Jing Xiong, and all other colleagues and friends, especially those working at 416 McLaughlin, who have created a very pleasant and warm environment.

Finally, I do not need say much about my parents and my wife, Ruozhu, for their selfless and endless love, support, and encouragement, which are more than words can describe…

This work is jointly supported by the University of California Transportation Center, the UC Berkeley Volvo Center for Future Urban Transport, the National Science Foundation, and the U.S. Department of Transportation (through the Dwight David Eisenhower Transportation Fellowship program).

Chapter 1

Introduction

1.1 Motivation

Promoting the use of public transit can be an effective way to reduce urban traffic congestion, fossil fuel depletion, and greenhouse gas emissions. In many cities of the world, however, the service quality of transit, especially bus systems, is very poor, which discourages travelers from choosing this more sustainable mode. A major cause of this poor service quality is that buses often experience severe congestion at busy bus stops (see Fig. 1.1 for an example).

Fig. 1.1 Bus congestion at a busy stop

While serving passengers at a busy stop, buses can interact in ways that limit bus discharge flow from the stop and create large bus delays. This can significantly degrade the bus system's overall service quality (Fernandez, 2010; Fernandez and Planzer, 2002; Gibson et al., 1989; TRB, 2003). Moreover, the disruptive bus maneuvers at these stops will often impede adjacent car traffic, and thus create more congestion. Hence, an important question in bus system planning is: how should one design a busy bus stop and manage bus operations in and around the stop to increase the rate that buses can discharge from the stop, and to mitigate the induced bus and car congestion?

To this end, models are needed to unveil the cause-and-effect relationships between the inputs for a busy stop (e.g., bus inflow, number of berths, etc.) and the outputs (e.g., the resulting bus delays). These models will help practitioners in selecting stop designs and operating strategies.

The research problem is explained in the next section. A review and critique of the relevant literature is furnished in Section 1.3. Contributions of this research are summarized in Section 1.4. The structure of the remaining chapters of this dissertation is presented in Section 1.5.

1.2 Research Problem

This dissertation is focused primarily on isolated bus stops where bus operations are not affected by neighboring stops and traffic signals. We further assume that buses enjoy an exclusive lane at least at the proximity of a stop (thus bus operations are not influenced by car traffic); and that sufficient space exists for storing the bus queues that can form immediately upstream of a stop.

The research problem includes two parts: (i) to formulate models to predict the service level (e.g., in terms of bus delay) of a bus stop as a function of a given bus inflow, Q, and key operating features, which will be explained below; and (ii) to apply these models to identify the most advantageous bus-stop designs and operating strategies. Part (i) serves as a step toward part (ii), and is the larger focus of this dissertation.

Metrics of service level for bus stops include the widely used failure rate, FR, defined as the probability that an arriving bus is temporarily blocked from using the stop by another bus (TRB, 2000; St. Jacques and Levinson, 1997); and average bus delay, $\bar{W}$, defined as the average time a bus spends at the stop and the queue upstream minus the time for loading and unloading passengers. We call the period of loading and unloading the bus's "service time."

To identify the operating features that have key influence on a bus stop's level of service, note that a stop operates like most *queueing systems*; see Fig. 1.2 for an example of a curbside stop. It exhibits a random pattern of bus (i.e., customer) arrivals and a distribution of bus service times. For simplicity, we assume in this dissertation that bus service times and arrival headways (i.e., inter-arrival times) are both independent, identically distributed (i.i.d.) random variables. Other key operating features include the number of berths (i.e., servers) in the stop, c, and the queue discipline, which pertains to whether a bus can overtake other buses downstream in and around the bus stop.

In this dissertation we examine two queue disciplines that are common in the real world, which we term "no-overtaking" and "limited-overtaking." At curbside stops, bus overtaking maneuvers are often prohibited both when a bus in the queueing area attempts to enter an empty berth, and when a bus that has finished serving passengers attempts to exit the stop. Cities often enact this "no-overtaking" regulation since an overtaking bus can disrupt car traffic in the adjacent lane(s). In other instances, overtaking maneuvers are allowed for buses when they attempt to exit the stop. This happens when buses can use the adjacent lane around curbside stops, and when buses dwell at off-line bus bays to serve passengers. These stops are termed “limited-overtaking” stops.[1]

Hence, the queueing systems can be denoted as GI/GI/c, as per Kendall’s notation (1953), but with an idling discipline (i.e., customers may be in queue while servers idle, because of the mutual blocking between the customers). In this sense, FR is the probability of delay in queue, and $\bar{W}$ is the expected wait time.

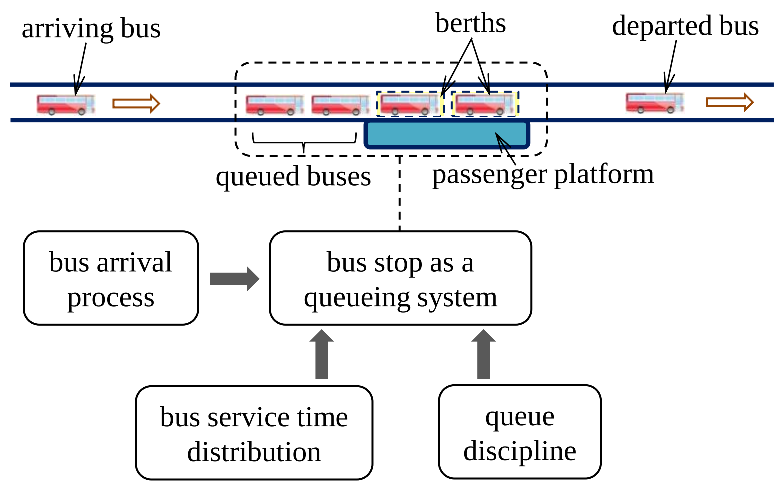


Fig. 1.2 A (curbside) bus stop as a queueing system

In light of all the above, the output of part (i) of the research problem can be expressed as follows:

$$\mathcal{L} = \Phi(Q, c, \mathcal{P}_H, \mathcal{P}_S, \mathcal{D}) , \tag{1.1}$$

where $\mathcal{L}$ denotes the service level metric (FR or $\bar{W}$); Φ denotes a function of the bus flow, Q, the number of berths, c, and other parameters defined next; the symbols $\mathcal{P}_H$ and $\mathcal{P}_S$ denote the probability distributions of bus arrival headways and service times, respectively; and $\mathcal{D}$ denotes the queue discipline.

In some cases we are more interested in the inverse problem of the above, expressed as:

$$Q = \Phi^{-1}(\mathcal{L}, c, \mathcal{P}_H, \mathcal{P}_S, \mathcal{D}) ; \tag{1.2}$$

i.e., to predict the maximum allowable bus flow, Q, as a function of the service level targeted by the transit agency, $\mathcal{L}$, and other operating features. This inverse problem is

[1] Since it is rather difficult to maneuver a bus’s tail into the berth, we assume that a queued bus can never enter an empty berth by overtaking other buses that are dwelling at the stop.

especially useful when determining the maximum bus flow that can be served by a stop given some target $\mathcal{L}$, and when identifying the capacity bottleneck in a congested bus system.

The models developed in this work are applied to (i) determine the suitable queue discipline and the minimum number of berths needed to serve a target bus flow given key inputs $(\mathcal{L}, \mathcal{P}_H, \mathcal{P}_S)$ of a bus stop; and (ii) evaluate the effects of select bus operating strategies. An example of these operating strategies is "stop splitting," in which a stop with multiple berths is partitioned into multiple neighboring stops.

1.3 Literature Review

This section summarizes the literature relevant to bus queueing at stops, including: the analytical recipes in professional handbooks (Section 1.3.1); findings in archival journals (Section 1.3.2); and relevant literature on methodologies for solving queueing systems that are similar to bus stops (Section 1.3.3).

1.3.1 Handbooks on Isolated Bus Stops

In the *Highway Capacity Manual*, or *HCM* (TRB 1985, 2000), the maximum allowable bus flow (termed "capacity" in the *HCM*) of an isolated single-berth stop is the inverse of the average bus service time adjusted by a reduction factor. The latter accounts for variations in bus arrival and service times, as well as the target service level. The reduction factor either has tabulated values based on the stop's designated failure rate (see Equation 12-10 and Table 12-17 of the 1985 *HCM*), or involves the standard Z-score corresponding to the failure rate and the coefficient of variation of bus service times (see Equation 27-5 of the 2000 *HCM*)[2]. With this reduction factor, the allowable flow of an isolated single-berth stop increases as failure rate increases, but only to a point. Curiously, the *HCM* formulas estimate the allowable bus flow to be maximal when the failure rate reaches 0.5. Intuition, on the other hand, tells us that this flow is maximum only when the failure rate is 1, such that a bus queue always exists. This intuition is confirmed in the dissertation; see Chapter 3. Additionally, the present work has unveiled a determinant of the allowable flow that has previously gone unreported: variation in bus headways to the stop.[3]

For multi-berth stops, the *HCM* calculates the allowable flow as the allowable rate of a single berth multiplied by a so-called "effective number of berths." The latter relies

[2] The standard Z-score is the score representing the area under one tail of the standard normal curve beyond the failure rate (see http://en.wikipedia.org/wiki/Standard_score for further discussion). The coefficient of variation is the ratio of the standard deviation to the mean.
[3] Although an earlier edition of the *HCM* includes a multiplicative adjustment factor that reportedly accounts for variations in bus arrival headway, the factor seems instead to account for FR (see Equation 12-7 and Table 12-17 of TRB, 1985).

again on tabulated values (see Table 12-19 of the 1985 *HCM* or Table 27-12 of the 2000 *HCM*). For each new berth added at a stop, the tables offer a corresponding effective number, depending on the queue discipline (i.e., whether overtaking permitted or not). Accordingly, each additional berth is presumed to yield a steadily diminishing gain in allowable flow, starting with the inclusion of a second berth at the stop. This consistent reduction in the marginal return appears to be an effort to account for the disruptive interactions that arise between buses at multi-berth stops.[4] Yet we find this is not true. Note that as an extreme case, when buses at berths never interfere with each other (the discipline is nonidling, i.e., a standard GI/GI/c queue), doubling the number of berths more than doubles the allowable flow (Whitt, 1993).[5] Also, the *HCM* tables furnish only one set of effective numbers for each queue discipline; i.e., the efficiencies lost in the presence of multiple berths are assumed to be independent of any other operating consideration, including the target service level, and the distributions of bus arrival and service times. This is at odds with our findings. Moreover, the tabulated values in the *HCM* show that limited-overtaking stops (i.e. the so-called "off-line" stops in the *HCM*) with a given number of berths always furnish higher allowable bus flows than do no-overtaking stops (i.e. the "on-line" stops in the *HCM*; see again Table 27-12 of the 2000 *HCM*). We also found that this is not correct (see Chapter 2).

Though the model forms and values of the tabulated factors changed slightly from one edition of the *HCM* to the next, both editions share the same fundamental ideas described above. These ideas are now part of the *Transit Capacity & Quality of Service Manual* (TRB 2003, Part IV, Chapter 1), which itself has been published with the intent of supplanting any discussion of transit systems in future editions of the *HCM*. *Transportation Planning Handbook* (ITE 1992, 1999) also adopts the same ideas.

1.3.2 Academic Literature on Isolated Bus Stops

Earlier critiques of the *HCM* formulas cited above appear in the literature. Gibson et al. (1989) pointed out that bus behavior at a congested stop follows a "complex stochastic process," and therefore criticized the method in the 1985 *HCM* to be of little use for most applications. Fernandez and Planzer (2002) tested the 2000 *HCM* formulas against field measurements and found that the formulas underestimated the stop's allowable flow. Both of the above references resorted to microscopic simulation because of the system's inherent complexity, and did not pursue analytical work.

Regarding queue disciplines, it is widely believed that all else equal, a limited-overtaking stop always provides higher allowable bus flow than a no-overtaking stop (e.g., Papacostas, 1982; Gardner et al., 1991). The works cited above used simulation and empirically-based models to support this belief. Due to the limitations of their methods, those works failed to analyze a wide range of operating conditions. Thus their conclusion turns out to be incomplete.

---

[4] This seems to be a consensus among researchers, including Gibson et al. (1989).
[5] This is because of the pooling effect. Details on this are furnished in Section 3.4.1.

There are also studies on how operating strategies can reduce bus queueing and improve allowable flows at stops. For example, stop splitting (i.e., splitting a stop with multiple berths into multiple neighboring stops) has been much studied via simulation (Gibson et al. 1989; St. Jacques and Levinson 1997; etc.). Most findings show that a stop with greater than 3 berths is generally unfavorable; though until now there have been no analytical methods for exploring this.

Finally, there is empirical evidence showing that bus arrivals at a multi-route bus stop often follow Poisson processes (Danas 1980; Kohler 1991; and Ge 2006), though Newell (1982) and Fernandez (2001) showed that this is not always the case. For bus service times, the Erlang distribution is found to fit well (Ge 2006). St. Jacques and Levinson (1997) suggests that a value of 0.4-0.8 be used as the coefficient of variation of these service times. Though oftentimes incomplete, these works provide useful guidelines on selecting suitable inputs to be used in our models.

1.3.3 Methodologies for Solving Queueing Systems

The dissertation relies to a large degree on means used to solve queueing systems similar to bus stops. These means include simulation and analytical methods (including approximations).

As described in the previous section, most studies of bus queueing at busy stops rely on computer simulation. To their credit, these simulation-based studies averted difficult analytical derivations, and still revealed useful insights. The insights, however, are limited because simulation models function as “black boxes,” as compared against analytical formulas that can directly unveil how inputs (e.g., the number of berths) influence outputs, such as bus delays. And to achieve a suitable degree of accuracy, simulation needs to be repeated numerous times to reduce the systematic and random errors generated in the simulation process. This will often mean high computational costs. Thus, most simulation-based efforts by necessity have focused on a relatively small number of select cases. This limitation might lead to incorrect or incomplete findings (e.g., Papacostas, 1982).

Nevertheless, simulation is useful, and is used in this research to verify the analytical results, and to solve certain queueing systems that are too complex to solve analytically. We develop analytical models (including approximations) whenever possible. Methods in queueing theory are used to this end.

Explicit closed-form solutions have been developed early in the 20th century for well-known queueing systems with Poisson customer arrivals (Erlang, 1909; Pollaczek, 1930; Khintchine, 1932). More sophisticated approximation models for queueing systems with general arrival/service patterns were developed later; see Whitt (1993) for a review of these models. However, none of the above-cited models can be used for bus stops. This is because the above works studied queueing systems where the servers are laid-out in parallel, such that a customer (e.g. a bus) can use any empty server (i.e. a bus berth) without being blocked by customers that are occupying other servers. As explained in Section 1.2, bus berths are laid-out in serial fashion along the stop, such that buses in the stop and the queue upstream can potentially block each other. This mutual blockage will inevitably increase the average bus delay, compared to a queueing system with the same number of servers, but laid-out in parallel. Thus, the analytical works cited above can only furnish lower bounds of the average bus delays for bus stops.

Regrettably, analytical models in the literature for queueing systems with serial servers are only applicable to simple cases of two servers where overtaking is prohibited and a customer queue is always present (Estrada et al., 2011; Gu et al., 2012; Hall and Daganzo, 1983). General analytical solutions to this queueing system do not exist in the present literature, to the best of our knowledge.

Still, mathematical methods and properties from the queueing theory literature can be used to solve the bus queueing problem at bus stops. These methods and properties may include the Poisson Arrivals See Time Averages (PASTA) property (Wolff, 1982), the method of embedded Markov chains (e.g. Neuts, 1978), the generating function method (Crommelin, 1932), the method of supplementary variables (Cox, 1955), etc. Some of these methods and properties are used in this dissertation.

1.4 Summary of Research Contributions

The previous section points to the need to develop a comprehensive set of analytical and simulation models that describe the unique operating features of serial bus berths in an isolated stop. To this end, we start by examining a simple, extreme case when a bus queue is always present upstream of the stop, i.e., when $\overline{W} = \infty$ and $FR = 1$. We develop analytical models in this case for a range of berth numbers, and for both "no-overtaking" and "limited-overtaking" stops. The models produce the maximum rates that buses can discharge from stops. These rates are compared to determine whether allowing bus overtaking is favorable under different operating conditions. The rates are also upper bounds for the flows of the general case, i.e., when $\overline{W} < \infty$ and $FR < 1$.

We then develop analytical and simulation models for the general case. These models can be used to predict bus delays (or allowable flows) for various operating conditions, and to select the suitable number of berths to serve a given bus flow. The models also shed light on how to choose the appropriate operating strategies for improving the bus service level at congested stops. These strategies include means to coordinate passenger boarding processes.

As a contribution to the queueing theory literature, the analytical methods developed for bus stops can be applied to other serial queueing systems, e.g., taxi queues, toll stations with tandem booths, and Personal Rapid Transit systems.

1.5 Organization of the Dissertation

The dissertation is organized as follows.

Chapter 2 presents the analytical solutions for the extreme case when $\bar{W} = \infty$ and $FR = 1$, for both "no-overtaking" and "limited-overtaking" stops. These two types of stop operations are then compared to identify which is more productive (in terms of allowable bus flow) for different operating conditions. A general control strategy that can potentially maximize the utilization of berths is also discussed.

Chapter 3 studies the general case when $\bar{W} < \infty$ and $FR < 1$ for a select type of stop operation: no-overtaking. We first develop analytical models for both single-berth and multi-berth stops of this type; for the latter we feature a Markov chain embedded in the stochastic bus operations at a stop. Parametric analysis is then presented using these models. The analysis furnishes insights into the cause-and-effect relationship between bus queueing and the operating features such as berth number.

Finally, Chapter 4 concludes the dissertation by summarizing its contributions, and discussing potential research opportunities that build upon the present work.

Chapter 2

The Extreme Case when $\bar{W} = \infty$ and $FR = 1$

In this chapter we examine the bus stops operating in an extreme case where a bus queue is always present upstream of the stop. In this extreme case, the bus discharge flow from the bus stop will reach its maximum. The service level, however, is the worst: $\bar{W} = \infty$ and $FR = 1$. Nonetheless, this extreme case is worth exploring, not only because it provides an upper bound of the allowable bus flow for the general case (i.e., when $\bar{W} < \infty$ and $FR < 1$), but also for its simplicity. To see why, note that in this extreme case the bus arrival process will have no influence on the bus discharge flow since buses all arrive at the rear of the queue. Thus Equation (1.2) reduces to:

$$Q_{extr} = {\Phi_0}^{-1}(c, \mathcal{P}_S, \mathcal{D}) \,, \tag{2.1}$$

where Q_{Extr} denotes the allowable bus flow in this extreme case; i.e., this flow only depends on the number of berths, c, bus service time distribution, and the queue discipline. The derivation of this simplified equation offers a simpler way to investigate problems such as comparing no-overtaking and limited-overtaking stops. The analysis of this extreme case furnishes other insights as well. These insights, to greater or lesser degrees, still hold in the general case, which is confirmed by the general-case models developed in Chapter 3.

This extreme case is especially simple when $c = 1$: both no-overtaking and limited-overtaking stops have the same allowable flow, $1/\bar{S}$, where $\bar{S}$ is the average bus service time. Thus, in this chapter we will explore the cases where $c \geq 2$.

We first develop the analytical models for no-overtaking and for limited-overtaking stops in Sections 2.1 and 2.2, respectively. A comparison between these two queue disciplines is furnished at the end of Section 2.2. Section 2.3 presents a general form of control strategy that can potentially maximize the use of serial berths. Section 2.4 summarizes the findings.

2.1 No-Overtaking Bus Stops

When buses are not allowed to make any overtaking maneuvers, queued buses will enter the stop in platoons of size c, and the time required to serve a platoon is the maximal bus service time across the platoon. The rate that buses discharge from the no-overtaking stop, Q_{NOT_extr}, is therefore:

$$Q_{NOT\_extr} = \frac{c}{\int_{t=0}^{\infty}(1-(F_S(t))^c)dt}, \tag{2.2}$$

where $F_S(t)$ is the cumulative distribution function (CDF) of the individual bus service time. The derivation of (2.2) is furnished in Appendix B.1.

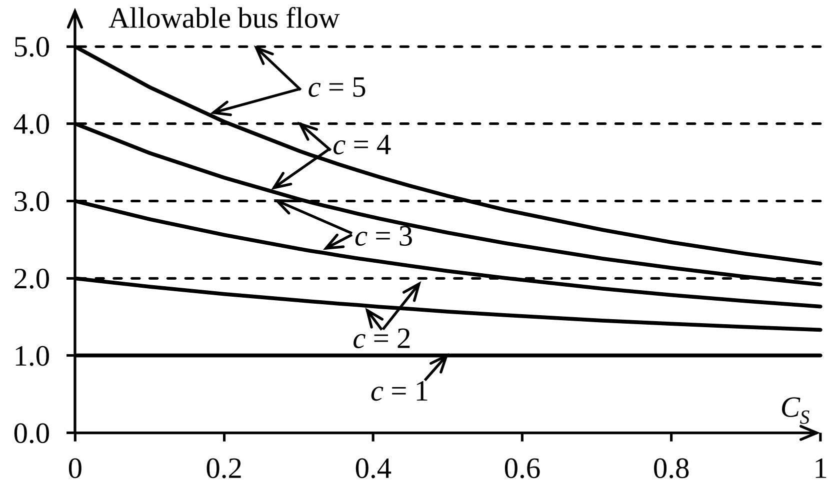
From (2.2), we plot in Fig. 2.1(a) the allowable flows (the solid curves) for $c = 1$-5 and $C_S = 0$-1, where C_S denotes the coefficient of variation in bus service time. We assume that bus service times follow gamma distributions with a fixed $\bar{S}$ (normalized as 1), since the gamma distribution is a generalization of the Erlang distribution, and the latter has been observed by Ge (2006) to be suitable for modeling bus service times.

Fig. 2.1(a) shows that the allowable flows decrease as C_S increases, save for the case of $c = 1$.[6] For comparison, the flows when the berths are assumed to be parallel (i.e., when buses can freely enter any empty berth or exit from any berth) are shown as the dashed lines. Comparisons between solid curves and dashed lines indicate that a large C_S significantly decreases the stop's capacity to serve buses. For example, when $C_S = 1$, the allowable flows produced by a 2-berth stop and a 3-berth stop are only 1.33 and 1.63 times that of a single-berth stop. This is also consistent with intuition, because a larger variation in bus service time means that buses are more likely to block each other in and around the stop.

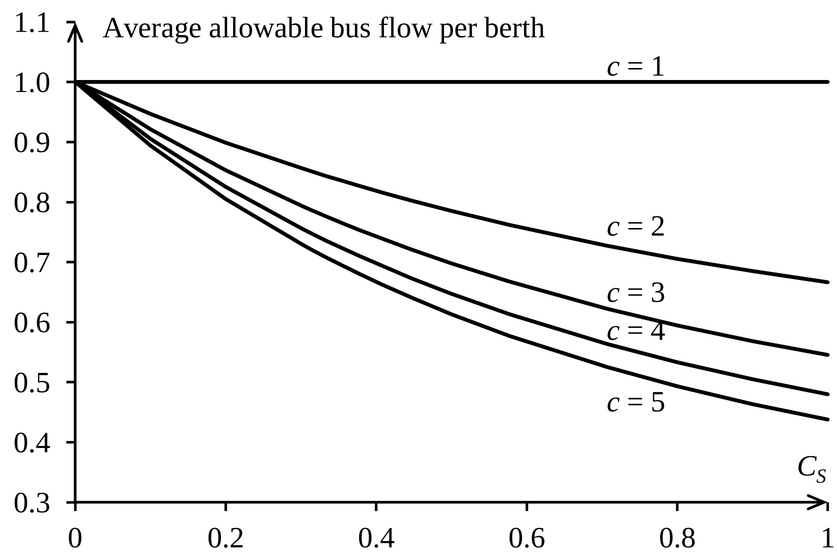
To further understand the effect of adding berths to a stop, we plot the average allowable flows per berth versus C_S in Fig. 2.1(b). It shows that the average allowable flow produced by each berth decreases as c increases. This is consistent with intuition because a larger c increases the possibility that buses block each other at the stop. This can also be seen from (2.2), by noting that the average allowable flow per berth, $1/\int_{t=0}^{\infty}(1-(F_S(t))^c)dt$, decreases with added berths, since the denominator increases with c.

---

[6] This is an analytical fact: Suppose X_i and Y_i are two sequences of gamma-distributed random variables (bus service time) with the same mean value, but X_i has a larger C_S than Y_i, then X_i is larger than Y_i in the convex sense. This in turn implies that $E[max(X_1, \dots, X_c)] \geq E[max(Y_1, \dots, Y_c)]$; i.e., the expected service time of a platoon of c buses with individual service times $X_1, \dots, X_c$ is larger than that with individual service times $Y_1, \dots, Y_c$.



(a) Allowable bus flows versus C_S



(b) Average allowable flows versus C_S

Fig. 2.1 Allowable flows and average allowable flows per berth of no-overtaking stops

2.2 Limited-Overtaking Bus Stops

At limited-overtaking stops, buses are allowed to overtake other buses dwelling at a downstream berth to exit the stop, but a queued bus cannot overtake any downstream bus to enter an empty berth. Analysis of stops of this type is more complicated than for no-overtaking stops. We will start by developing the analytical model for a 2-berth limited-overtaking stop (Section 2.2.1), and then discuss how to extend the analytical method to stops with more berths in a recursive fashion (Section 2.2.2).

2.2.1 Two-Berth Limited-Overtaking Stops

The analytical solution for no-overtaking stops presented in Section 2.1 exploits the property that buses are served in platoons. These bus platoons are separated by time instants when all the berths in the stop are empty. A similar property exists for limited-overtaking stops, as described next.

Imagine that at a certain point in time, both berths of a 2-berth limited-overtaking stop are empty. Two buses will then enter the stop in platoon. If the bus dwelling at the upstream berth finishes serving passengers first and leaves (by overtaking the downstream bus), a new bus from the upstream queue will fill the empty berth. But if the bus dwelling at the downstream berth exits when the upstream berth is occupied, no queued bus can fill the empty berth immediately, and that berth will remain empty until the upstream berth is also vacated. After that, another platoon of two queued buses will immediately enter the stop. Thus, we see that the time horizon is also divided into periods by time instants when the berths are both empty. We define these periods as *cycles*. Cycles of limited-overtaking stops differ from no-overtaking ones in that in the former case, buses in each cycle are not served in platoon, and the number of these buses is not fixed (recall that the number is c for no-overtaking stops). These differences greatly complicate the problem.

From the renewal theory, we have:

$$\begin{aligned} Q_{LOT\_extr} &= \lim_{t\to\infty} \frac{\text{number of buses served in } t}{t} \\ &= \frac{E[\text{number of buses served in a cycle}]}{E[\text{cycle time}]} \\ &= \frac{\bar{N}}{\bar{T}}, \end{aligned} \tag{2.3}$$

where $\bar{N}$ is the average number of buses served in a cycle, and $\bar{T}$ is the average duration of a cycle.

We denote S_j $(j = 1,2, \dots, N)$ as the service time of the j-th bus that enters the stop in a cycle, where N is the number of buses served in the cycle. Note from the above analysis that the cycle length, T, is equal to the time duration that the upstream berth is occupied, i.e.,

$$T = \sum_{j=2}^{N} S_j \, .$$

Recall that only one bus dwells at the downstream berth per cycle. Thus,

$$\begin{aligned} \bar{T} &= E\left[\sum_{j=1}^{N} S_j\right] - E[S_1] = E[N] \cdot E[S_j] - E[S_1] \\ &= (\bar{N} - 1)\bar{S} \, . \end{aligned}$$

Since we normalize $\bar{S}$ as 1, we have

$$Q_{LOT\_extr}(c=2)=\frac{\bar{N}}{\bar{N}-1}\,. \tag{2.4}$$

The next step is to calculate $\bar{N}$. Note that $N \geq 2$ for all cycles; if $S_1 > S_2$, then $N \geq 3$; if $S_1 > S_2 + S_3$, then $N \geq 4$; …. In light of this, we have

$$\bar{N}=2+\sum_{i=1}^{\infty} Pr\{S_1 > \sum_{j=2}^{i+1} S_j\}\,. \tag{2.5}$$

Given the distribution of bus service time, the second term on the right-hand-side of (2.5) can be calculated using Monte Carlo simulation. Alternatively, if the S_j's follow an Erlang-k distribution, then

$$Pr\left\{S_1 > \sum_{j=2}^{i+1} S_j\right\} = 2^{1-(i+1)k}\cdot\sum_{j=ik}^{(i+1)k-1}\binom{(i+1)k-1}{j}\,, \tag{2.6}$$

where $\binom{n}{m}=\frac{n!}{m!(n-m)!}$; a proof of (2.6) is furnished in Appendix B.2. Combining (2.4-2.6), we obtain:

$$Q_{LOT\_extr}(c=2)=1+1/\left(1+\sum_{i=1}^{\infty}2^{1-(i+1)k}\cdot\sum_{j=ik}^{(i+1)k-1}\binom{(i+1)k-1}{j}\right)\,. \tag{2.7}$$

This can be further simplified to (see Appendix B.3):

$$Q_{LOT\_extr}(c=2)=1+1/\left(1+\sum_{j=1}^{k}\frac{2\cdot(-1)^{k-j}}{(k-j)!}\cdot\left.\frac{d^{k-j}\left(\frac{x^{-(j+1)}}{x^{k}-1}\right)}{dx^{k-j}}\right|_{x=2}\right)\,. \tag{2.8}$$

The right-hand-side of (2.8) can be computed numerically, using differences to approximate the differentiations, i.e., $\left.\frac{dy}{dx}\right|_{x=x_0}\approx\frac{y(x_0+\delta)-y(x_0-\delta)}{2\delta}$, where δ is a small number.

Using the above equations, we plot in Fig. 2.2 the average allowable flow per berth of a 2-berth limited-overtaking stop for $C_S \in [0,1]$ (see the dash-dot curve). We assume Erlang-distributed bus service times, but for other distributions we obtain similar results. The figure shows that the average allowable bus flow still decreases as C_S increases.

The average allowable flow for a 2-berth no-overtaking stop is also plotted for comparison (see the solid curve). Surprisingly, the comparison shows that allowing buses to perform overtaking maneuvers can increase the bus discharge flow *only* when $C_S > C_{S,crit} = 0.41$. For example, the allowable flow is increased by 12.5% when $C_S = 1$. However, when C_S is low, permitting overtaking actually diminishes the bus discharge flow, e.g., by 17% when C_S is close (but not exactly equal) to 0.[7] This is at odds with conventional wisdom (e.g., Papacostas, 1982; Gardner et al., 1991). The present finding can be explained by the damaging behavior of "greedy bus drivers" as follows.

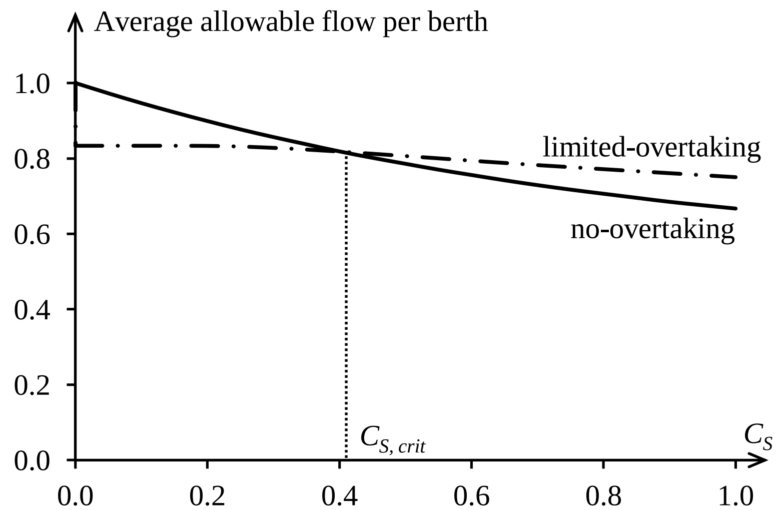


Fig. 2.2 Comparison of limited- and no-overtaking stops with 2 berths

Consider the cycle described in the beginning of this section that starts with two buses entering the stop in platoon style. If the bus dwelling at the upstream berth leaves first, the vacated berth will be immediately occupied by a "greedy" bus from the upstream queue. Now if the bus dwelling downstream departs the stop soon after, the downstream berth will be left empty for a relatively long while until the "greedy" bus in the upstream berth finishes its service. This effect will cause a reduction in bus discharge flow. Further note that this negative effect of "greedy drivers" becomes dominant when C_S is low, i.e., when the service times of the first two buses are similar, because in this case it is more likely that the third bus will enter the upstream berth only moments before the downstream-most dwelling bus pulls out of the stop. On the other side, the negative effect diminishes as C_S increases, and then the limited-overtaking stops become superior to no-overtaking stops.

Hence, the choice of queue discipline depends on the variation in bus service time, as Fig. 2.2 shows. Bus overtaking maneuvers should be prohibited when C_S is low, but might be allowed otherwise. The next section confirms this finding for bus stops with more than 2 berths.

[7] If $C_S = 0$ (i.e., when the bus service time is constant), bus operations at limited- and no-overtaking stops are identical. Thus, the dash-dot curve in Fig. 2.2 has a jump at $C_S = 0$.

2.2.2 Limited-Overtaking Stops with More than Two Berths

The analytical models for limited-overtaking stops when $c > 2$ can be developed in a recursive fashion; i.e., the model of c-berth stops can be built upon the results of $(c-1)$-berth stops. For simplicity, in this section we only present the model for 3-berth stops.

As before, we define a cycle as the period between two consecutive time instants when all the berths in the stop are empty. Recall that (2.3) still holds:

$$Q_{LOT\_extr}(c=3) = \frac{\bar{N}(c=3)}{\bar{T}(c=3)} .$$

We further define a *subcycle* as the period between two consecutive time instants when the two upstream berths of the stop are both empty. Note that a cycle may include one or more subcycles. Moreover, the number of buses served in a subcycle and the subcycle length are both statistically equivalent to those of a cycle for a 2-berth limited-overtaking stop. Thus, we have

$$\bar{N}(c=3) = 1 + \bar{N}_{SC} \cdot \bar{N}(c=2)$$

and

$$\bar{T}(c=3) = \bar{N}_{SC} \cdot \bar{T}(c=2) ,$$

where $\bar{N}_{SC}$ denotes the average number of subcycles in a cycle. Hence, we have

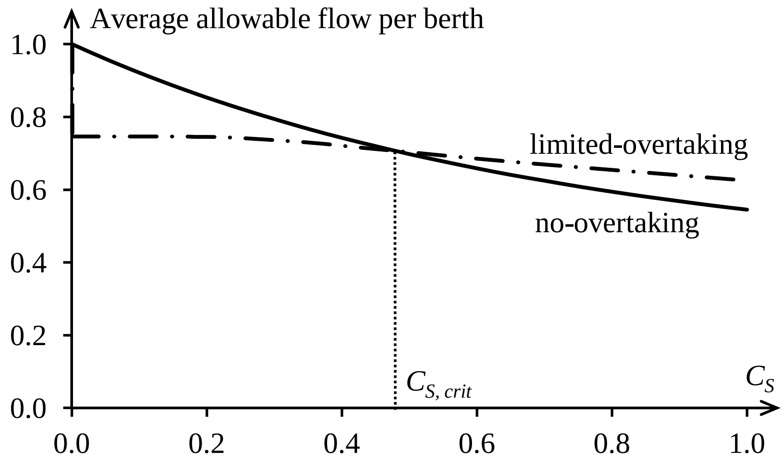
$$Q_{LOT\_extr}(c=3) = \frac{1+\bar{N}_{SC}\cdot\bar{N}(c=2)}{\bar{N}_{SC}\cdot(\bar{N}(c=2)-1)} . \qquad (2.9)$$

$\bar{N}_{SC}$ can be calculated in a way similar to $\bar{N}(c=2)$ (see (2.5)):

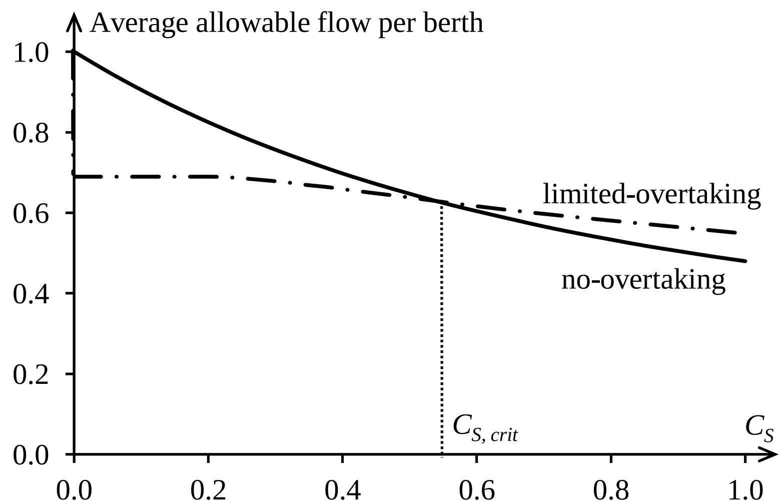
$$\bar{N}_{SC} = 1 + \sum_{i=1}^{\infty} Pr\{S_1 > \sum_{j=1}^{i} T_{SC,j}\} , \qquad (2.10)$$

where $T_{SC,j}$ denotes the duration of the j-th subcycle in a cycle. The probabilities on the right-hand-side of (2.10) can be computed either by Monte Carlo simulation, or by integrating over the distribution functions of S_1 and $T_{SC,j}$. The distribution function of $T_{SC,j}$ is given in Appendix B.4.

In theory, the above method can be applied recursively to calculate the allowable flows for limited-overtaking stops with any number of berths. But the computational complexity increases rapidly as c increases. In this dissertation we only show the results for the allowable flows of 3-berth and 4-berth stops. These results are shown as dash-dot curves in Figs. 2.3(a) and (b) for $C_S \in [0,1]$. We still assume Erlang-distributed bus service times. Average allowable flows for no-overtaking stops are also shown for comparison (the solid curves).



(a) $c = 3$



(b) $c = 4$

Fig. 2.3 Comparison of limited- and no-overtaking stops with 3 and 4 berths

The patterns in Figs. 2.3(a) and (b) are similar to Fig. 2.2: the allowable flows are decreasing in C_S; limited-overtaking stops produce higher allowable flows for higher C_S (i.e., $C_S > C_{S,crit} = 0.47$ for $c = 3$, and $C_S > C_{S,crit} = 0.54$ for $c = 4$), and no-overtaking stops are better otherwise. Compared to Fig. 2.2, Figs. 2.3 (a) and (b) also show that having more berths means a larger detrimental effect of "greedy drivers," and a greater threshold of C_S for allowing overtaking.

The negative effect of "greedy drivers" can possibly be reduced by controlling the damaging behavior of those drivers. This possibility is explored next.

2.3 A General Strategy that Controls the Use of Berths

We first note that a simple control is to prohibit overtaking for $C_S \leq C_{S,crit}$, and allow overtaking otherwise. This simple strategy will produce the upper envelope of the two curves shown in Fig. 2.2 for 2-berth stops, or in Figs. 2.3(a) or (b) for 3- or 4-berth stops.

A more general form of control can be used to further improve the allowable bus flow. Taking a 2-berth stop as an example, this general control will be applied at the time instant when the upstream berth is vacated (and thus a queued bus can enter the berth), but the downstream berth is still occupied. We call this instant a *control point*. The idea is to allow the next queued bus to enter the emptied upstream berth at a control point (and thereby potentially block other buses from entering the downstream berth later in the cycle) only when the remaining service time estimated at the downstream berth is sufficiently long. This remaining service time could be readily estimated using: i) the elapsed service time at the downstream berth at the control point; and ii) the bus service time distribution. To see how this general control strategy can improve the allowable bus flow, we develop a near-optimal form of this control strategy that can approximately maximize the utilization of berths. The work will be presented for 2-berth (limited-overtaking) stops. Brief discussion about stops with more than 2 berths will come at the end of this section.

We will first develop the condition under which it is advantageous to block any queued bus from entering the upstream berth at a control point. To this end, we denote t_E and t_R as the elapsed and the remaining service times of the downstream-most dwelling bus at the control point; and n_0 as the number of buses that have been served or that are being served in the present cycle. Further denote (n_1, t_1) as the expected number of buses served in the cycle and the cycle duration when a decision is made to block the next queued bus from entering the (empty) upstream berth; and (n_2, t_2) if the opposite decision is made. The (n_1, t_1) and (n_2, t_2) are functions of n_0 and t_E. Finally we assume that a cycle with measurements n_0 and t_E occurs with probability p; and that the average bus discharge flow for the rest of the cycles is $q = \bar{n}/\bar{t}$, where $\bar{n}$ and $\bar{t}$ are the average number of buses served per cycle and the average cycle length for the rest of the cycles, respectively.

With the variables defined above, we find that it is opportune to block the next bus from entering the stop if

$$\frac{(1-p)\bar{n}+pn_1}{(1-p)\bar{t}+pt_1} \geq \frac{(1-p)\bar{n}+pn_2}{(1-p)\bar{t}+pt_2}.$$

This is equivalent to:

$$(1-p)(\bar{n}t_2 + n_1\bar{t}) + pn_1t_2 \geq (1-p)(\bar{n}t_1 + n_2\bar{t}) + pn_2t_1 .$$

For continuously distributed bus service times, p is generally close to 0. Thus we ignore the second terms on both sides of the above inequality, and have:

$n_1\bar{t} - \bar{n}t_1 \geq n_2\bar{t} - \bar{n}t_2$.

Dividing both sides above by $\bar{t}$ and moving the right hand side to the left, we have the condition for blocking the next bus:

$$\Delta n = (n_1 - qt_1) - (n_2 - qt_2) \geq 0 \ . \tag{2.11}$$

Note that $(n_1 - qt_1)$ and $(n_2 - qt_2)$ are the extra numbers of buses that would be served in the cycle (compared to the average level) when the two control decisions are made. Thus, Δn represents the additional number of buses served if the next bus is blocked, as compared to the other decision.

From (2.11) we see that the value of q is needed to make the control decision. When $p \to 0$, q is approximately the allowable bus flow of the stop after applying control, and is therefore unknown. However, the simple form of control described in the beginning of this section provides a lower bound of q. For simplicity, we will use this lower bound as an approximation of q. This approximation may compromise the effect of control.[8] Yet when the approximation is close to the real value of q, the results will be a good approximation.

We now develop n_1, t_1, n_2, and t_2 as functions of n_0 and t_E. For the case when the next queued bus is blocked,

$$n_1 = n_0 \ , \tag{2.12}$$
$$t_1 = E[t_E + t_R] = E[S_1 | S_1 \geq t_E] \ . \tag{2.13}$$

The other case is more complicated, because the calculation should include not only the bus that is allowed to enter at the present control point, but also the buses that may be allowed to enter the stop at any possible control points later in the cycle. Thus n_2 and t_2 shall be determined in a recursive way. To avoid this complexity, we ignore any further buses that might be served later in the present cycle, except for the current bus in question. In real cases, the possibility of having two or more control points in a cycle is small, thus the approximation should be fairly accurate. Thus, we have:

$$n_2 \approx n_0 + 1 \ , \tag{2.14}$$
$$t_2 \approx E[t_E + max\{t_R, S'\}] = E[max\{S_1, t_E + S'\} | S_1 \geq t_E] \ ; \tag{2.15}$$

where S' denotes the service time of the bus that is allowed to enter at the control point.

Combining (2.11-15), we obtain the following condition:

[8] Using (2.11), we can infer that if q is underestimated, the control strategy will tend to choose the decision with a larger cycle length, i.e., to allow a few moderately damaging buses to enter the stop.

$$\Delta n = q \cdot \frac{\int_{s=0}^{\infty}(1-F_S(s))F_S(t_E+s)ds - F_S(t_E)}{1-F_S(t_E)} - 1\ . \qquad (2.16)$$

Note that this condition is independent of n_0. The derivation of (2.16) is shown in Appendix B.5.

For a given bus service time distribution, (2.16) can be used to numerically calculate the range of t_E suitable for blocking the next queued bus. Note that when the bus service time is exponentially distributed, (2.16) reduces to $q \geq 2$, which is a false condition. It means that queued buses should be allowed to enter the stop whenever the upstream berth is vacated. This is consistent with intuition because when the service time is memoryless, any remaining service time will be statistically equivalent to a new service time.

We find that if bus service times follow the Erlang-k distribution, the range of t_E suitable for blocking the next bus can be expressed as $[t_{E\_crit}, \infty]$, where t_{E_crit} denotes the critical elapsed service time. This is also aligned with intuition. We plot t_{E_crit} as a function of C_S in Fig. 2.4. This figure shows that t_{E_crit} increases as C_S increases; i.e., the strategy tends to allow more buses to enter the upstream berth at control points when the bus service time is more unpredictable. It also shows that when C_S is very large, t_{E_crit} approaches infinity; i.e., one should always choose to allow the bus's entry. On the other side, $t_{E\_crit} \rightarrow 0.53$ when $C_S \rightarrow 0$, and thus the probability of $t_E < t_{E\_crit}$ is very low since in this case bus service times are almost always near 1; i.e., one would choose to block the bus's entry to the upstream berth at almost all control points.

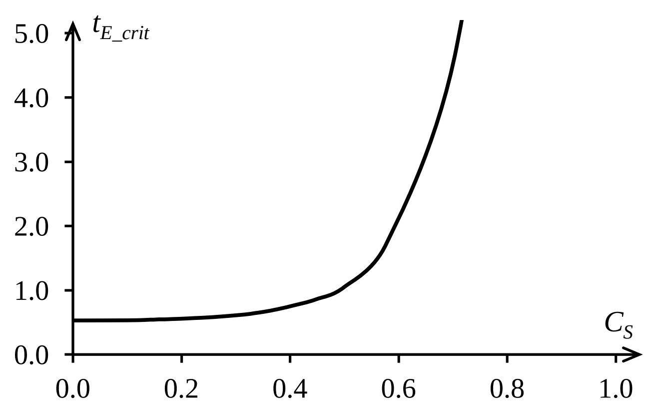


Fig. 2.4 t_{E_crit} versus C_S

We develop a simulation model to estimate the allowable bus flow for a 2-berth limited-overtaking stop when the control strategy is applied. The algorithm used in the simulation model is furnished in Appendix C.1. The estimated average allowable flows are then plotted in Fig. 2.5 against C_S (the double-line solid curve).

The allowable flows for no-overtaking (the solid curve) and limited-overtaking stops (the dash-dot curve) are also shown in Fig. 2.5 for comparison. The comparison unveils that the allowable flow curve produced by applying the general control almost overlaps the upper envelope of no- and limited-overtaking curves. Although the general control can further improve the allowable flow when $C_S \approx 0.4$, the improvements are less than 2%. This can be explained in part by the facts that: (i) when C_S is low, the general control is not beneficial since in this case one should almost always choose to block the bus; and (ii) when C_S is moderate or high, the general control is not accurate because t_E can hardly provide accurate estimates of t_1 and t_2 (see again (2.13) and (2.15)). Thus, the simple form of control described in the beginning of this section is arguably sufficient and is easy to implement.

The effects of the general control strategy can be analyzed in a similar way for stops with more than 2 berths. However, due to the same reasons as explained above, we believe that the improvements provided by the general control would still be modest.

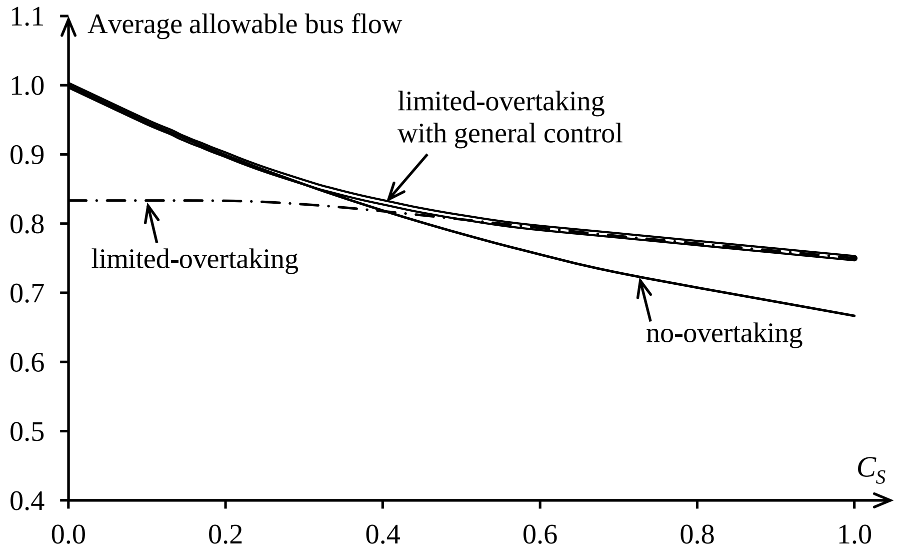


Fig. 2.5 Allowable flows with and without general control versus C_S

2.4 Summary of Findings

This chapter explored the allowable bus flows of the two types of bus stops – no-overtaking and limited-overtaking stops – in the extreme case where a bus queue is always present. Analytical models were developed and from these we obtained upper bounds of the allowable bus flows for the general case when the bus queue is finite. The models unveil many insights, including:

i) The allowable bus flow decreases when the variation in bus service time increases.

ii) The average allowable bus flow produced by each berth decreases as more berths are added to the stop.

iii) For a given number of berths, limited-overtaking stops are not always better than no-overtaking ones. Limited-overtaking stops produce higher bus flow when the variation in bus service time is large, but less flow when the variation is small. This finding can be explained by the damaging behavior of "greedy drivers."

iv) A simple control strategy can be applied based on the value of C_S. Taking a 2-berth stop as an example, the strategy prohibits bus overtaking maneuvers when $C_S < 0.4$, and allows these maneuvers otherwise. This simple control is shown to be sufficient when compared to a more complicated, general control strategy.

These insights have practical implications. For example, the transportation authority probably would choose to prohibit bus overtaking maneuvers that are disruptive to adjacent car traffic, and under certain operating conditions this overtaking prohibition would benefit both cars and buses. The authority can also improve the bus discharge flow from an extremely busy stop by means of reducing C_S. One possible means to do this is to organize the passenger boarding process at the stop.

Most insights presented above are obtained by assuming Erlang- or Gamma-distributed bus service times. Yet the methodologies developed in this chapter can be applied to any service time distribution. And we have found that most results are similar for other distinct distributions. We understand that some results might be different if "special" distributions are used. For example, if bus service times follow a Bernoulli distribution, then the remaining service time of the bus dwelling at the downstream berth (of a 2-berth stop) can sometimes be accurately predicted. Thus, the general control strategy might bring considerable benefits in this case. However, service time distributions like this are very rare in the real world.

Most of the above insights hold in the general case when the bus queue is finite. This is confirmed in the next chapter.

Chapter 3

The General Case when $\bar{W} < \infty$ and $FR < 1$

In this chapter we explore the bus operations at stops for the general case, i.e., when $\bar{W} < \infty$ and $FR < 1$. In this case the bus inflow never exceeds the allowable bus flow for the extreme case obtained in Chapter 2. The analysis of this general case is important because it is the case that concerns practitioners. For example, a transit agency is often interested in the maximum bus discharge flow from a stop for a target value of $\bar{W}$ or FR, or in means to reduce bus delays at a congested stop. To this end, one needs to understand the relations between the bus flow and the stop's service level. Moreover, the bus arrival process at the stop also has a significant effect on the stop's service level (as we shall see in the present chapter), which was ignored by the extreme-case models developed in the previous chapter.

In light of this, we develop analytical and simulation models for bus stops of the general case, and use these models to examine i) the returns in allowable bus flow by adding berths to the stop; and ii) the effects of bus arrival process and service time distribution on the stop's allowable bus flow, Q. For simplicity, the discussion is focused on no-overtaking stops. Simulation has confirmed that most findings from stops of this kind hold for limited-overtaking stops as well.

We start by developing analytical models for the simplest case, i.e., when $c = 1$ (Section 3.1). More general analytical models for multi-berth stops are presented in Section 3.2. Using these models, we compare the two service level metrics, $\bar{W}$ and FR, and show that the former is more suitable for use (Section 3.3). The select service-level metric, $\bar{W}$, is then used for the analyses of the returns in allowable bus flow (Section 3.4) and to explore the effects of bus arrival process and service time distribution (Section 3.5). Section 3.6 summarizes the findings of this chapter.

3.1 Single-Berth Stops

It will be assumed that bus stops operate in steady state, such that the bus arrival process and the service time distribution are both time invariant. In steady state, the average bus inflow to the stop always equals the average outflow, so the overall inflow rate must be less than the maximal service rate.

To simplify our analysis and highlight the findings, we start by assuming two special cases in regard to the bus arrival process: Poisson arrivals (in Section 3.1.1), as can occur when the stop serves multiple bus routes;[9] and uniform bus arrivals[10] (in Section 3.1.2), as may occur, at least in theory, when the stop serves a single route with buses that are rigidly controlled. These two queueing systems are denoted as M/G/1 and D/G/1, respectively, as per the notation of Kendall (1953). Finally, Section 3.1.3 examines the case of a more general bus arrival pattern (GI/G/1). Recall that the bus arrival headways are assumed to be independent of the bus service times.

3.1.1 Poisson Bus Arrivals – M/G/1

In steady state, Poisson bus arrivals to a stop satisfy the PASTA (Poisson Arrivals See Time Averages) property; see Wolff (1982). This implies that FR is equal to the fraction of time that the stop's single berth is utilized. Note that this utilization fraction is equal to the bus flow, Q, since the mean service time is normalized to 1:

$$FR = Q. \tag{3.1}$$

As per intuition, (3.1) shows that the single-berth's allowable bus flow is maximal when $FR = 1$ (recall that this is at odds with the *HCM* formulas). It further shows that for Poisson bus arrivals and FR as the service level metric, the allowable bus flow is independent of the variation in bus service time. This independence turns out not to hold, however, if $\bar{W}$ is used as the service level metric.

The formula for $\bar{W}$ can be obtained by applying the well-known Pollaczek-Khinchine formula (Pollaczek, 1930; Khinchine, 1932):

$$\bar{W} = \frac{Q(1+{C_S}^2)}{2(1-Q)} \ . \tag{3.2}$$

Equation (3.2) shows that $\bar{W}$ approaches infinity as Q approaches 1. This can be seen from the solid curves in Fig. 3.1 for $C_S = 0$, 0.5, and 1. Fig. 3.1 also shows that for a given target $\bar{W}$, Q decreases as C_S increases. This is consistent with the finding from the extreme case. Recall that the finding is not true when using FR in the general case.

[9] For simplicity, in this dissertation we make the assumption of Poisson bus arrivals for both single-berth (Section 3.1.1) and multi-berth (Section 3.2) stops. However, we note that this assumption might be valid only when the bus flow is low, as per Newell (1982).

[10] Here "uniform bus arrivals" mean deterministic, evenly-spaced bus arrivals. This is different from bus arrivals that are uniformly distributed. We will use uniform distribution to model bus service times in Section 3.2.5.

3.1.2 Uniform Bus Arrivals – D/E_k/1

Assume now that the bus arrival headways are deterministic and equal, denoted as D in Kendall's notation; and that bus service time follows an Erlang-k distribution, denoted as E_k. For this present case, our model does not have a closed-form solution. An analytical model that can be solved numerically is derived in Appendix B.6.

The resulting allowable bus flows are shown against FR and $\bar{W}$ with solid curves in Figs. 3.2(a) and (b), respectively, for $C_S = 0$, 0.1, 0.5, and 1. These curves collectively reveal that, for uniform bus arrivals, Q increases as the coefficient of variation in bus service time diminishes, no matter which service level metric is chosen. The upper limit of Q, however, is independent of C_S. The case of $C_S = 0$ corresponds to the perfect coordination of bus arrivals and bus service time, such that $FR = \bar{W} = 0$. The curves in this idealized case therefore reduce to a point, also as shown in Figs. 3.2(a) and (b).

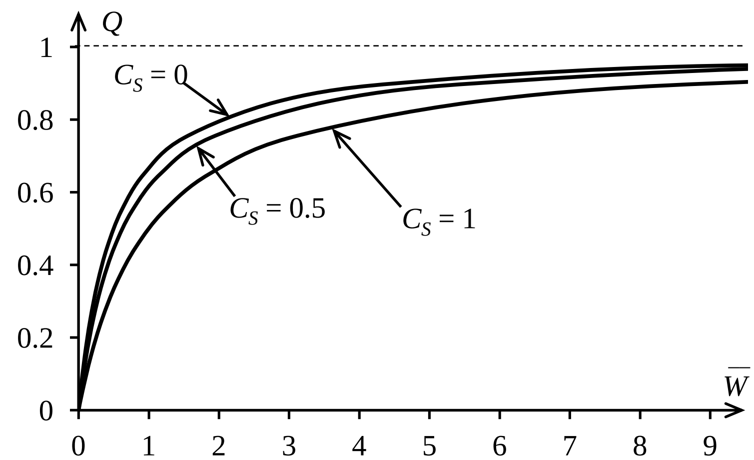


Fig. 3.1 Q versus $\bar{W}$ for single-berth stops with Poisson bus arrivals

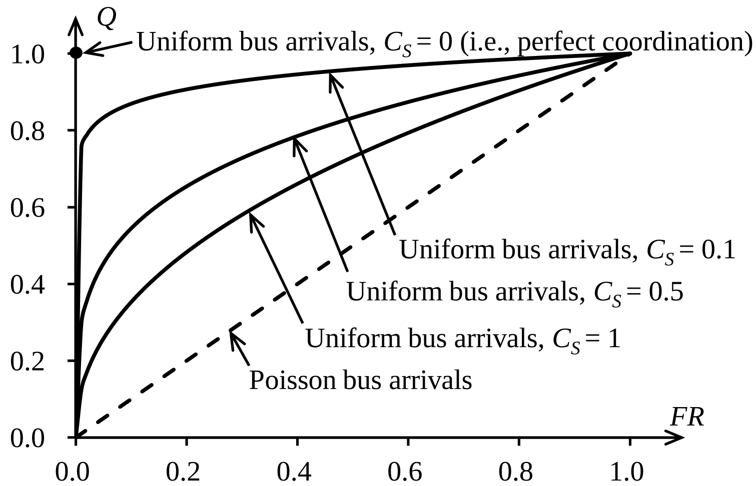
The relations for Poisson bus arrivals revealed in (3.1) and (3.2) are shown in Figs. 3.2(a) and (b) as well; see the dashed curves. Comparing the dashed curves against the solid ones reveals that Q also increases with diminishing variation in bus headways. (We can see this because the coefficient of variation is 0 and 1 for uniform and Poisson bus arrivals, respectively).

3.1.3 General Bus Arrivals – E_j/E_k/1

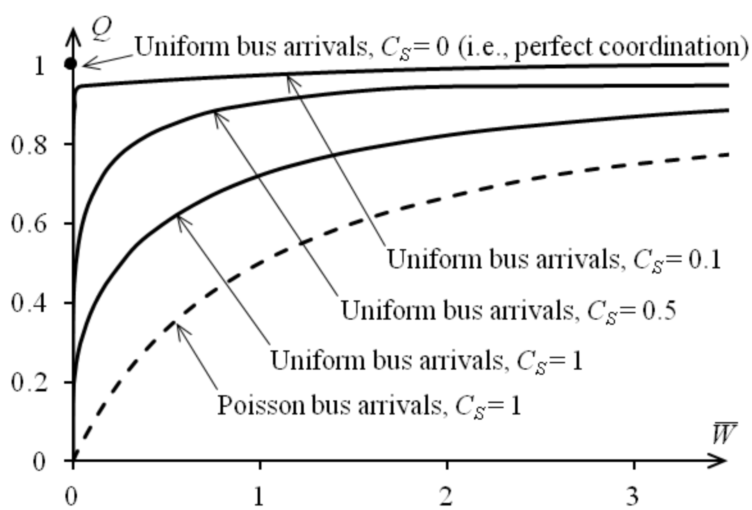
We continue to model bus service time as above, and now use the Erlang-j distribution to describe a more general distribution for bus headways. A numerical solution was derived in similar fashion to the uniform bus-arrival case.

The results confirm that Q decreases as C_S or C_H increases, where C_H is the coefficient of variation of bus arrival headways. The effects of C_H can be seen from Fig

3.3, where the allowable bus flows are plotted against $\bar{W}$ for $C_S = 0.5$ and $C_H = 0$ (uniform arrivals), 0.5, and 1 (Poisson arrivals).



(a) Q versus FR



(b) Q versus $\bar{W}$

Fig. 3.2 Q versus FR and $\bar{W}$ for single-berth stops with uniform bus arrivals

3.2 Multi-Berth Stops – Analytical Solutions

In this section we develop analytical solutions for no-overtaking stops with an arbitrary number of berths. We assume that buses arrive at the stop as a Poisson process[11], and that the distribution of an individual bus's service time is independent of the stop's number of berths. With these assumptions, we identify a Markov chain embedded in the stochastic bus operations at the stop. This Markov chain is then used to develop exact

[11] As mentioned above, this assumption is realistic for low traffic scenario; see Footnote 9.

solutions for two special cases: a stop with deterministic bus service time and c berths, where c can be any positive integer; and a stop with a general service time distribution but only two berths. These are denoted as M/D/c/SRL and M/G/2/SRL systems, respectively, as per the notation of Kendall (1953); and we use SRL, an abbreviation of "serial", to denote the queue discipline at a no-overtaking stop. The generating function (i.e., z-transform) method is utilized to obtain these exact solutions. Finally, the exact solutions are used to formulate a closed-form approximation of Q as a function of the select service-level metric for the general case. The methodology can be applied with virtually any reasonable metric for service level. For simplicity, in this section we choose $\bar{W}$ as our metric, which considers both the bus delay suffered in the entry queue, and any extra dwell time encountered by the bus at the stop itself after having served passengers there.

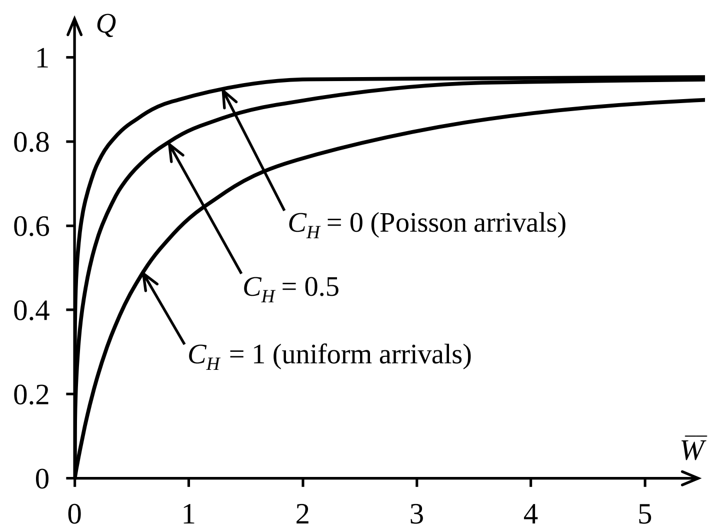


Fig. 3.3 Q versus $\bar{W}$ for single-berth stops with general bus arrivals and $C_S = 0.5$

The Markov chain is described in Section 3.2.1. The transition probabilities of the Markov chains for the two special cases are furnished in Section 3.2.2. The balance equations are then formulated and solved for the Markov chains' limiting probabilities (Section 3.2.3). These results are used to calculate the service level metrics for the two special cases (Section 3.2.4). The closed-form approximation is presented in Section 3.2.5. The discussion will emphasize the logic used for these matters. Details of the derivations are relegated to appendices.

3.2.1 The Embedded Markov Chain

Similar to the extreme case studied in Chapter 2, the general case is characterized by recurrent time instants when the buses in all berths have discharged from the stop. We define these time instants as *regenerative points*. Analogously, a *cycle* is defined as the time interval between two successive regenerative points. Note that a cycle of the general case differs from that of the extreme case in that in the former case, buses in each cycle are not served in platoon, and the number of these buses is not constant.

Let $\tilde{L}_n$ be the number of buses queued at the stop's entrance at the beginning of the n-th cycle (i.e., the n-th regenerative point); and recall that Q is the rate of (Poisson) bus arrivals, and c the stop's number of berths. We have the following result:

Claim: given Q, c, and the distribution of bus service times at the stop, the stochastic process $\{\tilde{L}_n\}$ is a Markov chain.

To see that the claim is true, note first that from the n-th regenerative point and thereafter, the bus arrival process, by virtue of being Poisson, is independent of any historical information. The service process is also independent of historical information, except for $\tilde{L}_n$, because at the regenerative point all berths are empty, and only $\tilde{L}_n$ and c determine the number of buses that enter the stop and the length of any residual bus queue immediately thereafter.

The bus-stop's service level (and other of its steady-state properties) can be obtained using the Markov chain once its transition probabilities are determined.

3.2.2 Transition Probabilities

Note that Q has a supremum (i.e. a minimum upper bound)[12] above which the system operates in an unstable state with infinite bus queues. So that the steady-state distribution of the embedded Markov chain is well defined, we assume Q is less than its supremum. We further define the Markov chain's transition probabilities:

$$P_{i,j} = Pr\{\tilde{L}_{n+1} = j | \tilde{L}_n = i\} .$$

For the special case of the M/D/c/SRL system, the $P_{i,j}$ are formulated as functions of Q and c, and details of this formulation are given in Appendix D.1. The $P_{i,j}$ for the M/G/2/SRL case are formulated as functions of Q and the CDF of bus service time, $F_S(t)$, as described in Appendix D.4.

3.2.3 Balance Equations of Limiting Probabilities and their Solutions

Let $\boldsymbol{P} = [P_{i,j}, i \geq 0, j \geq 0]$ be the matrix of transition probabilities; π_i $(i \geq 0)$ the limiting probability that the Markov chain is in state i, i.e. $\pi_i = \lim_{n\to\infty} Pr\{\tilde{L}_n = i\}$; and $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots]$ the limiting distribution of the Markov chain. Thus, $\boldsymbol{\pi}$ is the solution to the balance equation:

[12] This supremum is the allowable bus flow in the extreme case, as developed in Chapter 2.

$\boldsymbol{\pi} = \boldsymbol{\pi P}$.

Our solution method uses the z-transform of $\boldsymbol{\pi}$,
$\tilde{\pi}(z) = \sum_{i=0}^{\infty} \pi_i z^i$,

to consolidate the infinite-size balance equation into a single functional equation (Crommelin, 1932). Its solution can be converted (e.g. using the inverse z-transform) back to $\boldsymbol{\pi}$, the original distribution. Details on the solutions for the two special cases are given in Appendix D.2 and D.5.

3.2.4 Average Bus Delay

The above results are now used to estimate the average bus delay, $\bar{W}$. (Other service-level metrics could be obtained as well.) $\bar{W}$ is taken to be the sum of two normalized average delays: in the entry queue, $\bar{W}_q$, and in the berth after the bus has served its passengers, $\bar{W}_b$.

Determining $\bar{W}_q$ requires the calculation of the average number of buses in queue over time, $\bar{L}_q$, which is equal to the average of the queue length seen by each Poisson bus arrival, thanks to the PASTA property (Wolff, 1982). We obtain the $\bar{L}_q$ by taking the ratio of two averages: (i) the average of the sum of the queue lengths seen by each bus arrival in cycle n, TL_n (this average is denoted as $\overline{TL}$); and (ii) the average of the number of bus arrivals in that cycle, A_n (this average is denoted as $\bar{A}$). This is because:

$$
\begin{aligned}
\bar{L}_q &= \lim_{T\to\infty} \frac{\text{sum of queue lengths seen by all arrivals during } T}{\text{number of arrivals during } T} \\
&= \lim_{T\to\infty} \frac{\text{sum of queue lengths seen by all arrivals during } T/\text{number of cycles during } T}{\text{number of arrivals during } T/\text{number of cycles during } T} \\
&= \frac{\lim_{T\to\infty} \frac{\text{sum of queue lengths seen by all arrivals during } T}{\text{number of cycles during } T}}{\lim_{T\to\infty} \frac{\text{number of arrivals during } T}{\text{number of cycles during } T}} \\
&= \frac{\overline{TL}}{\bar{A}}\,.
\end{aligned}
$$

To obtain $\overline{TL}$ and $\bar{A}$, we consider the following four scenarios. These describe the possible states of the system at the start and end of each n-th cycle.

Scenario 1: No bus queues are present at either the start or the end of the n-th cycle (i.e. $\tilde{L}_n = \tilde{L}_{n+1} = 0$). In this scenario, no bus arriving during cycle n encounters a queue (i.e. $TL_n = 0$); and the number of buses that arrive during cycle n, A_n, is the number served during that cycle, M_n (i.e. $A_n = M_n$).

Scenario 2: A bus queue is present at the start of cycle n ($\tilde{L}_n = i > 0$), but no bus is present at the end of that cycle ($\tilde{L}_{n+1} = 0$). To satisfy $\tilde{L}_{n+1} = 0$, the i can be no greater than c; and $TL_n = 0$. Further note that $A_n = M_n - i$.

Scenario 3: A queue of size $i \leq c$ is present at the start of cycle n ($\tilde{L}_n = i \leq c$), and a queue persists at the end of that cycle ($\tilde{L}_{n+1} = j > 0$). In this scenario, the stop is filled during the cycle; i.e. $M_n = c$ and $A_n = c + j - i$. The first $c - i$ arrivals fill unused berths, such that the first $(c - i + 1)$ arrivals see no entry queue. The $j - 1$ arrivals to follow will see successively longer queues that range from 1 to $j - 1$. Thus,

$$TL_n = \frac{1}{2} j(j-1) .$$

Scenario 4: A queue size greater than c is present at the start of cycle n ($\tilde{L}_n = i \geq c + 1$), and a queue thus persists at the end of that cycle ($\tilde{L}_{n+1} = j \geq i - c > 0$). In this scenario, as in the previous one, $M_n = c$ and $A_n = c + j - i$. And since the earliest moments of cycle n are characterized by $i - c$ buses that remain in the entry queue, arrivals thereafter will see queue lengths in the sequence $i - c, i - c + 1, \ldots, j - 1$. Thus,

$$TL_n = \frac{1}{2}(i + j - c - 1)(c + j - i) .$$

Note from the above that, for known c, the TL_n and A_n depend only on $\tilde{L}_n$, $\tilde{L}_{n+1}$, and M_n. Thus, $\overline{TL}$ and $\bar{A}$ can be obtained by taking weighted averages:

$$\overline{TL} = \sum_{i,j,k} Pr\{\tilde{L}_n = i, \tilde{L}_{n+1} = j, M_n = k\} \cdot TL_n \text{ , and}$$
$$\bar{A} = \sum_{i,j,k} Pr\{\tilde{L}_n = i, \tilde{L}_{n+1} = j, M_n = k\} \cdot A_n \text{ ,}$$

where $Pr\{\tilde{L}_n = i, \tilde{L}_{n+1} = j, M_n = k\}$ is the long-run probability of a cycle where $\tilde{L}_n = i$, $\tilde{L}_{n+1} = j$, and $M_n = k$. This probability is equal to $\pi_i \cdot Pr\{\tilde{L}_{n+1} = j, M_n = k | \tilde{L}_n = i\}$; and the derivations of $Pr\{\tilde{L}_{n+1} = j, M_n = k | \tilde{L}_n = i\}$ for the two special cases are furnished in Appendix D.1 and D.4, respectively.

Therefore, we have:

$$\bar{L}_q = \frac{\sum_{i,j,k} Pr\{\tilde{L}_n = i, \tilde{L}_{n+1} = j, M_n = k\} \cdot TL_n}{\sum_{i,j,k} Pr\{\tilde{L}_n = i, \tilde{L}_{n+1} = j, M_n = k\} \cdot A_n} . \tag{3.3}$$

The average bus delay in the queue is then obtained from Little's formula (Little, 1961):

$$\bar{W}_q = \bar{L}_q / Q.$$

The above recipe for $\bar{W}_q$ is applied to M/D/c/SRL systems in Appendix D.3. Given that the service time is deterministic, the average bus delay in berths, $\bar{W}_b$, is 0 in this case.

For M/G/2/SRL systems, $\bar{W}_b$ can be obtained similarly to the above, as shown in Appendix D.6. This latter appendix also describes the determination of $\bar{W}_q$ for M/G/2/SRL systems.

3.2.5 An Approximation for General Service Time Distribution and Arbitrary c

The exact solutions for the M/D/c/SRL and M/G/2/SRL systems account for the effects on $\bar{W}$ from c and from C_S, respectively. Hence we can formulate an approximation for M/G/c/SRL systems by applying an idea similar to those previously used in the queueing literature (e.g. Maalφe, 1973; Nozaki and Ross, 1975):

$$\bar{W}_{M/G/c/SER}(\rho, c, C_S) \approx \frac{\bar{W}_{M/D/c/SER}(\rho,c)}{\bar{W}_{M/D/c/SER}(\rho,2)} \cdot \bar{W}_{M/G/2/SER}(\rho, C_S), \tag{3.4}$$

where the $\bar{W}$ denote the average bus delays for the queueing systems described by the subscripts; and the service ratio, ρ, is the ratio of bus inflow to the supremum of the bus discharge flow from the stop (i.e. the allowable bus flow of the extreme case). Thus, ρ takes a value between 0 and 1.

To find a closed-form approximation using (3.4), we first obtain approximations for $\bar{W}_{M/D/c/SER}$ and $\bar{W}_{M/G/2/SER}$ by fitting least squares models to the exact solutions. Both of these $\bar{W}$'s increase as ρ increases; they are 0 when $\rho = 0$, and they approach infinity as $\rho \to 1$. In light of this, we tried combinations of algebraic fractions and exponential functions, monomials, trigonometric functions, etc., as candidate forms of least squares models. From the many candidates tried, we found that the square errors of the following forms (with the corresponding parameters) are modest:

$$\bar{W}_{M/D/c/SER}(\rho, c) = \frac{0.19c+0.061}{c-0.54}\left(\tan\frac{\pi}{2}\rho\right)^{(-0.065c+1.21)} \tag{3.5}$$

and

$$\bar{W}_{M/G/2/SER}(\rho, C_S) = (0.29C_S + 0.29)\left(\tan\frac{\pi}{2}\rho\right)^{(0.046C_S+1.10)}, \tag{3.6}$$

where for M/G/2/SRL systems, we assume that bus service time is uniformly distributed.

Inserting (3.5) and (3.6) into (3.4), we obtain an approximation for average bus delay:

$$\bar{W}_{M/G/c/SER}(\rho, c, C_S) = \frac{0.63c+0.20}{c-0.54}(0.29C_S + 0.29)\left(\tan\frac{\pi}{2}\rho\right)^{(-0.065c+0.046C_S+1.23)}. \tag{3.7}$$

A formula is then derived from (3.7) to estimate the stop's allowable bus flow for a specified target average delay. To this end, we exploit a relationship between the allowable bus flow, Q, and ρ. For uniformly distributed bus service time, this relationship is

$$Q = \frac{c}{1+\sqrt{3}C_S\frac{c-1}{c+1}}\rho, \tag{3.8}$$

as derived in Appendix B.7.

Combining (3.7) and (3.8):

$$Q = \frac{2c}{\pi(1+\sqrt{3}C_S\frac{c-1}{c+1})}\tan^{-1}\left[\left(\frac{c-0.54}{(0.63c+0.20)(0.29C_S+0.29)}\bar{W}\right)^{\frac{1}{-0.065c+0.046C_S+1.23}}\right]. \quad (3.9)$$

If $\bar{S} \neq 1$, a bus-stop's allowable flow for a specified $\bar{W}$ can be estimated by $Q/\bar{S}$, where Q is determined by (3.9). If the target average bus delay is not normalized but instead has units (e.g. minutes), then $\bar{W} = \bar{w}/\bar{S}$, where $\bar{w}$ is the un-normalized target average bus delay.

It turns out that (3.9) is not only a good approximation for uniformly-distributed service times, but also works well for certain other service time distributions. As an illustration, the fitness of (3.9) is shown in Figs. 3.4(a) and (b) for $c = 3$ and 4, respectively. In each figure, the approximations (the dark solid curves) are compared with: (i) exact results obtained from the M/D/c/SRL model shown with the smaller, dark dashed curve for $C_S = 0$; (ii) simulated results for uniformly distributed service time (the longer, dark dashed curve) for $C_S = 0.4$;[13] and (iii) simulated results for gamma-distributed service time (the dark dotted curves) for $C_S = 0.4$ and 0.8. The thin, light solid curves will be explained momentarily. Visual inspection of Figs. 3.4(a) and (b) shows that the approximation fits the exact values very well except when C_S is large and service times are gamma distributed. Even in this latter case, the errors are still modest.

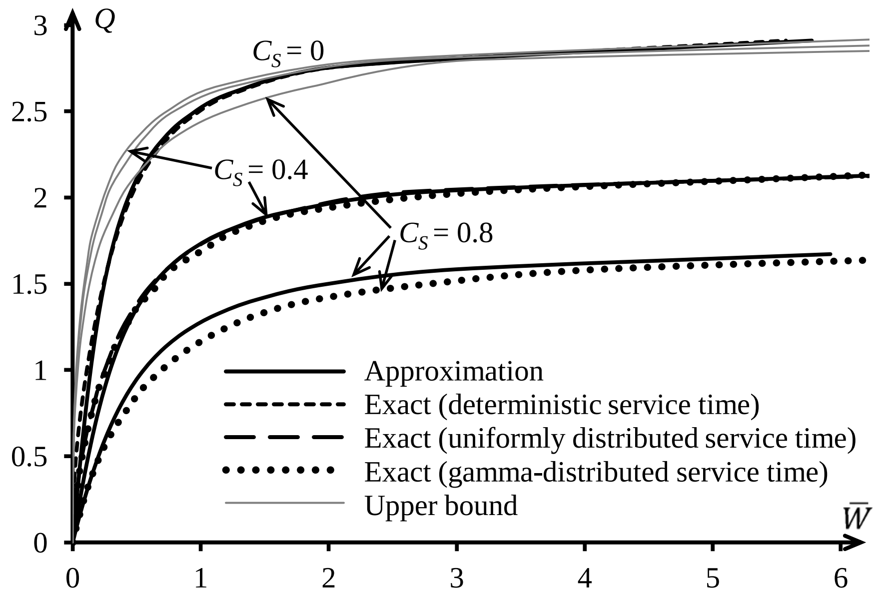
One might be interested in knowing how much of the reduction in the allowable bus flow is caused by buses blocking each other at the stop. To explore this matter, curves of allowable bus flow are plotted for some "ideal" stops where buses can freely enter any empty berths or exit the stop whenever they have finished serving passengers there; see the thin, light solid curves in Figs. 3.4(a) and (b). These curves were also produced by simulation. From both figures, we can see that the influence of mutual bus blocking is minimum when $C_S = 0$, and increases significantly as C_S increases.

3.3 Comparison of Service Level Metrics: FR versus $\bar{W}$

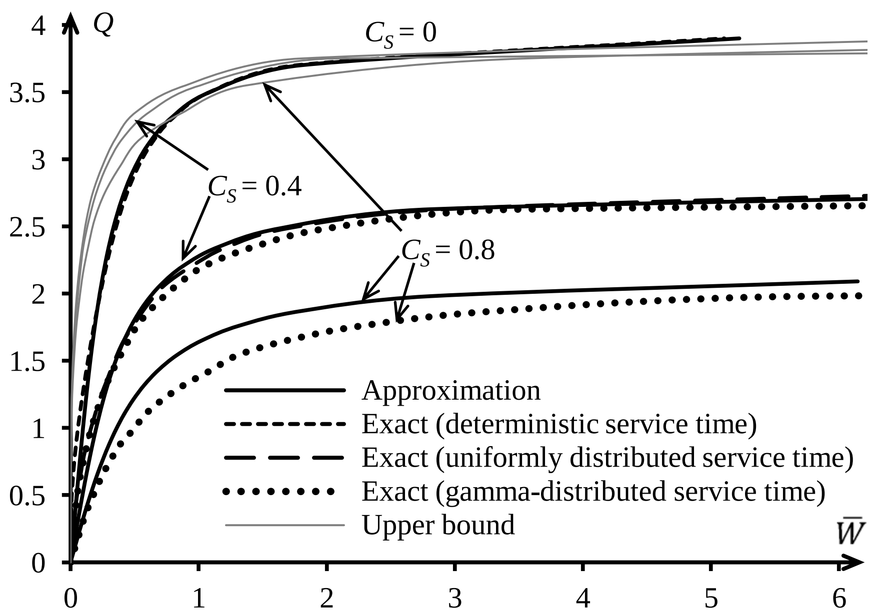
Average bus delay is the service level metric that both bus users and operators care about most, because it is readily observable, and obviously relevant to their concerns regarding travel times and operating costs. Failure rate, however, is the service-level metric most commonly used in the literature on bus stops (e.g. TRB, 2000). We presume that this is because FR is thought to be a good proxy of $\bar{W}$. We will now show that this assumption regarding the suitability of FR is not the case.

---

[13] Note that for uniformly distributed service time, the maximum C_S is $\sqrt{1/3}$, hence we do not show a longer, dark dashed curve for $C_S = 0.8$.



(a) $c = 3, C_S = 0, 0.4, 0.8$



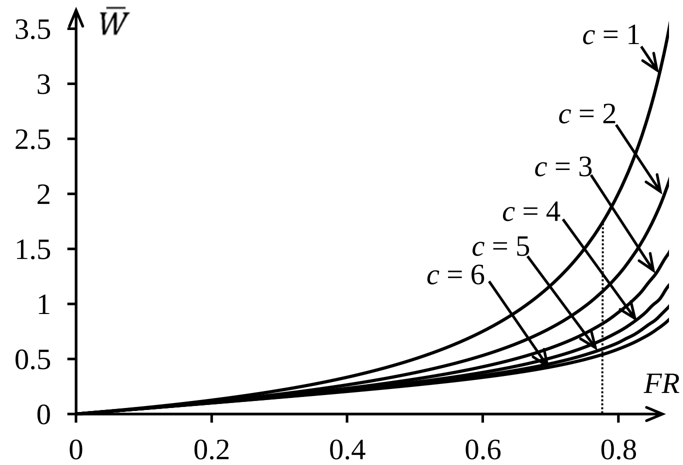
(b) $c = 4, C_S = 0, 0.4, 0.8$

Fig. 3.4 Fitness of the M/G/c/SRL approximation

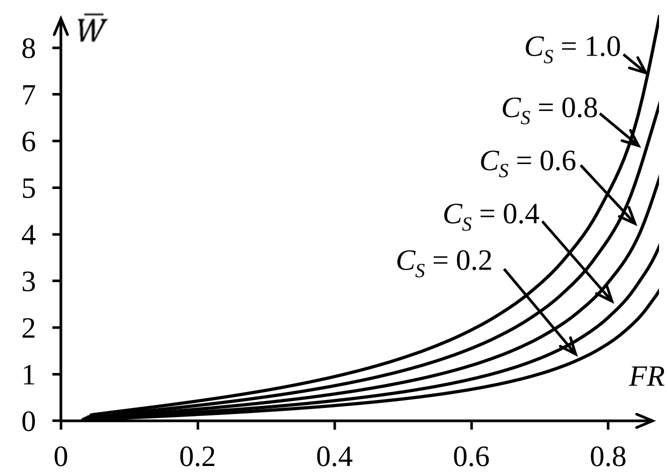
Our exact solutions are used to predict both FR and normalized (unitless) $\bar{W}$ for a range of operating environments. The two metrics are compared for M/D/c/SRL systems over a realistic range of c in Fig. 3.5(a). Comparisons for M/G/2/SRL systems are made assuming gamma-distributed service time and for a range of C_S in Fig. 3.5(b).

The relations shown in both figures are non-linear. The figures further reveal that, for small values of $\bar{W}$, modest changes in that metric coincide with disproportionally large changes in FR. Yet, FR is relatively insensitive to changes in $\bar{W}$ when $\bar{W}$ is large.

Further note that for different numbers of berths, the same FR corresponds to different values of $\bar{W}$; e.g., see the dashed vertical line in Fig. 3.5(a). This means that when adding berths to a stop, the average delay can increase significantly if the same FR is used as the service-level measure.



(a) M/D/c/SRL



(b) M/G/2/SRL

Fig. 3.5 Average bus delay versus failure rate

Finally, note that FR does not account for the average bus delay in berths, $\bar{W}_b$. (The importance of $\bar{W}_b$ will be discussed in the next section.) In light of the above, we will chose $\bar{W}$ as the service-level metric for the remaining analysis of the dissertation.

3.4 Returns in Allowable Bus Flow

In this section we examine how the bus stop's allowable bus flow changes by adding berths to the stop. Previous studies claimed that the allowable bus flow always diminishes when c increases (e.g., TRB, 2000; Gibson et al., 1989). However, as we will show in this section, the relation is in fact more complicated. To understand why, we start by introducing two competing effects that influence the allowable bus flow of multi-berth stops (Section 3.4.1). These effects are termed the "blocking" and the "berth pooling" effects. The returns in allowable bus flow from added berths are then studied for special cases that isolate the above effects in Sections 3.4.2 and 3.4.3, respectively. The general case is analyzed in Section 3.4.4.

3.4.1 Two Competing Effects

Discussion begins with the blocking effect. A bus can enter a stop only when its upstream-most berth is open. (At this time, the entering bus proceeds as far as possible until encountering the end of the stop or a dwelling bus; and the entering bus will then dwell at the downstream-most available berth for its entire time in the stop.) Similarly, a bus can discharge from a stop only after all buses that were previously dwelling at that stop's downstream berths have departed. Apparently, this blocking effect for entering and exiting a stop tends to diminish the stop's returns in Q brought by added berths. The effect diminishes, however, when the bus flow decreases.

We illustrate the second effect, berth pooling, with the following example. Consider two independent, single-berth stops, each with an equal bus arrival rate, λ, as shown on the left side of Fig. 3.6. (Dashed boxes in this figure denote berths, and shaded rectangles denote buses). If we ignore the blocking effect, the fluctuations in bus arrivals would be better served by pooling the two berths into a single, double-berth stop, as shown on the right side of Fig. 3.6. Thus for the same total bus arrival rate (2λ for both the left and right sides in the figure), this berth pooling effect means that the double-berth stop would enjoy a lower bus queueing delay than would the two single-berth stops; i.e., the double-berth stop would have a higher allowable bus flow for a given $\bar{W}$. Thus, berth pooling tends to improve the stop's returns in Q brought by added berths. The effect diminishes, however, when the bus flow increases and buses form long queues upstream of the stop (such that the fluctuations in bus arrivals become insignificant).

Thus, we see that the above effects act in opposing directions: for large Q's, the pooling effect diminishes while the blocking effect dominates;[14] while for small Q's, the opposite holds. Surprisingly, however, the blocking effect also dominates when $Q \rightarrow 0$, as will be explained later. We will therefore isolate the two effects by examining multi-berth stops under these special cases where one effect clearly dominates.

[14] An exception can occur under perfect coordination; i.e., when platoons of c buses arrive at uniform intervals and the service time is constant. In this case, neither blocking nor berth pooling take effect.

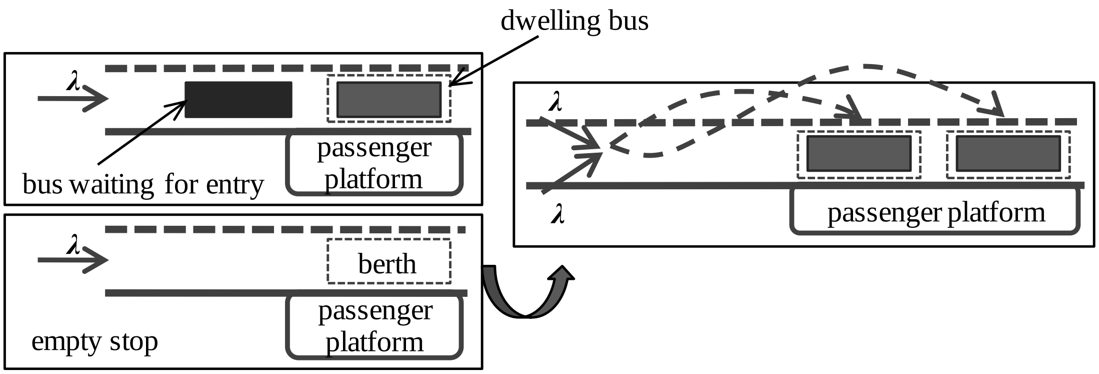


Fig. 3.6 Berth pooling effect

3.4.2 Returns in Q when Blocking Effect Dominates

As shown above, the blocking effect dominates when Q approaches its maximum, i.e., when a bus queue is always present upstream of the stop. This extreme case has been examined in Chapter 2; and as shown in Fig. 2.1(b), the returns in Q diminish due to the blocking effect.

We find as an exception that the blocking effect also dominates when $Q \to 0$. To understand why, recall that the service level metric, $\overline{W}$, consists of two components: the average bus delays in the entry queue, $\overline{W}_q$, and in the berth after the bus has served its passengers, $\overline{W}_b$. When $Q \to 0$, $\overline{W}_b$ dominates $\overline{W}$, and thus leads to decreasing returns in Q. This was to be expected: fewer berths can reduce the time that a bus spends at its berth awaiting other buses to depart from their own berths downstream.

As an illustration, Fig. 3.7 was constructed using the exact solution for M/G/2/SRL systems, and for a range of small $\overline{W}$ (i.e. $\overline{W} \leq 0.5$). Note from this figure how $\overline{W}_b$ accounts for increasing proportions of $\overline{W}$ as the latter decreases. Note too how this dominance increases with increasing C_S.

To illustrate how dominant $\overline{W}_b$ call for diminishing returns in Q from added berths, consider Fig. 3.8. It was constructed using the approximation for M/G/c/SRL systems by assuming $C_S = 0.2$ and $c =$ 2-5. The curves are shown with different line-types to make the comparison clear. Note from this figure how added berths bring decreasing returns in allowable bus flow, even for the low C_S. Further note that this phenomenon will be ignored when FR is used as the service-level metric, since FR does not account for $\overline{W}_b$.

Fig. 3.8 also shows that as $\overline{W}$ increases, the average allowable bus flows for $c = 3$ and $c = 4$ become greater than that for $c = 2$. This change will be explained next.

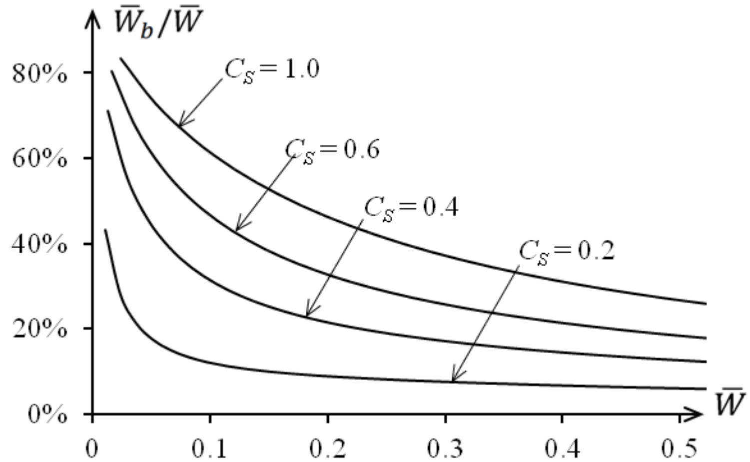


Fig. 3.7 $\bar{W}_b$ dominates $\bar{W}$ for low $\bar{W}$ ($c = 2$)

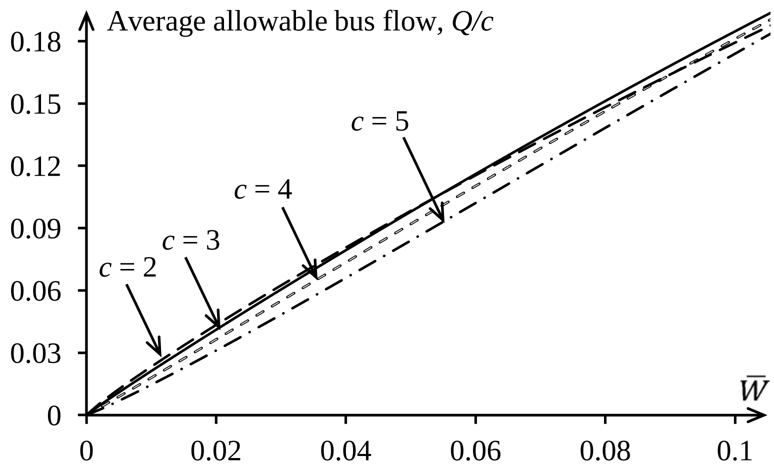


Fig. 3.8 Decreasing returns in allowable flow at very low $\bar{W}$ values ($C_S = 0.2$)

3.4.3 Returns in Q when Berth Pooling Effect Dominates

The $\bar{W}_b$ dominates $\bar{W}$ only when Q (and equivalently $\bar{W}$) is very low. As $\bar{W}$ increases, $\bar{W}_b$ becomes negligible as compared to $\bar{W}_q$ (see again Fig. 3.7). The berth pooling effect will therefore dominate. As an illustration, Fig. 3.9 shows the average allowable bus flow per berth for $C_S = 0.2$, $c = 2$-5 and $\bar{W} = 0.1$-0.6. It was again constructed using the approximation (3.9). Note from this figure how added berths bring increasing returns in allowable bus flow.

The increasing returns in Q are also found when using FR as the service-level metric. Note that this finding calls into question what handbooks (and the academic literature) have to say on the subject; i.e., the implication that added berths bring decreasing returns in allowable bus flow does not hold in general. More interesting evidence in this regard comes next.

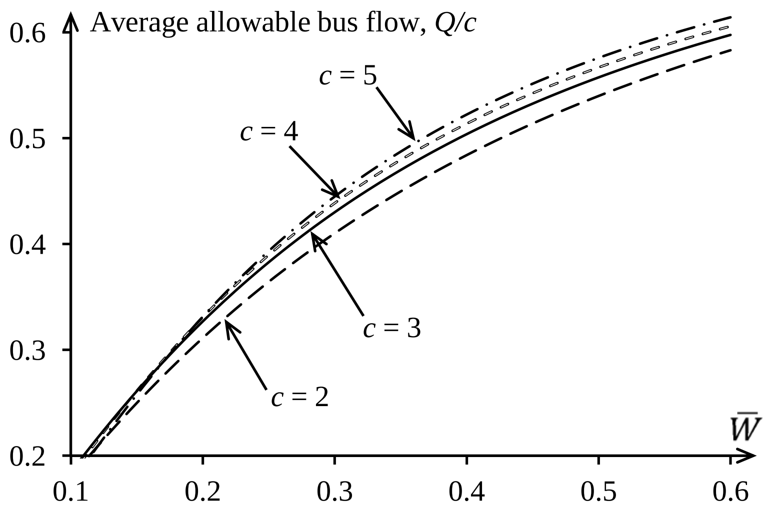


Fig. 3.9 Increasing returns in allowable flow at modest $\overline{W}$ values ($C_S = 0.2$)

3.4.4 Returns in Q in the General Case

We now show using Fig. 3.10 how the returns in Q vary when the $\overline{W}$ change from small to large values. We still use (3.9) for calculation and assume $C_S = 0.2$, $c = 1$-5. The curves reveal how the countervailing effects of blocking and berth pooling produce mixed results in terms of the allowable bus flow returned by adding berths to a stop.

When $\overline{W}$ is low (but not approaching 0), additional berths can produce increasing returns in Q, thanks to the berth pooling effect. Toward the other extreme, the curves reveal that added berths produce diminishing returns in Q. This is because the blocking effect tends to dominate. For intermediate $\overline{W}$ values, the results are found to be mixed in that the return in Q for an extra berth: increase for small c; and decrease for larger c. For example, at the vertical dashed line in Fig. 3.10, the curve for $c = 2$ lies above the curve for $c = 1$, meaning that adding a second berth brings increasing returns in Q. However, this favorable trend does not continue: note that the curve for $c = 3$ lies below the curve for $c = 2$.

These findings are logical in light of what was unveiled for the two competing effects in Sections 3.4.1-3. Yet, our finding that returns in Q vary with $\overline{W}$ runs counter to the *HCM*'s suggestion in this regard; i.e. using a single set of numbers for “effective berths” evidently does not suffice for all operating environments. (Note that similar findings can be obtained by using FR, the *HCM*'s choice for service level metric.)

Moreover, parametric analysis performed with our models indicates that returns in allowable flow are influenced by C_S, such that the case shown by Fig. 3.10 does not hold in general. For example, Fig. 3.11(a) presents our predictions for $C_S = 0$. Note that for the large range of $\overline{W}$ shown, returns in allowable flow increase, whatever the c. Fig.

3.11(b) shows that the opposite occurs when $C_S = 1$. More details regarding the effects of C_S will be presented in the next section.

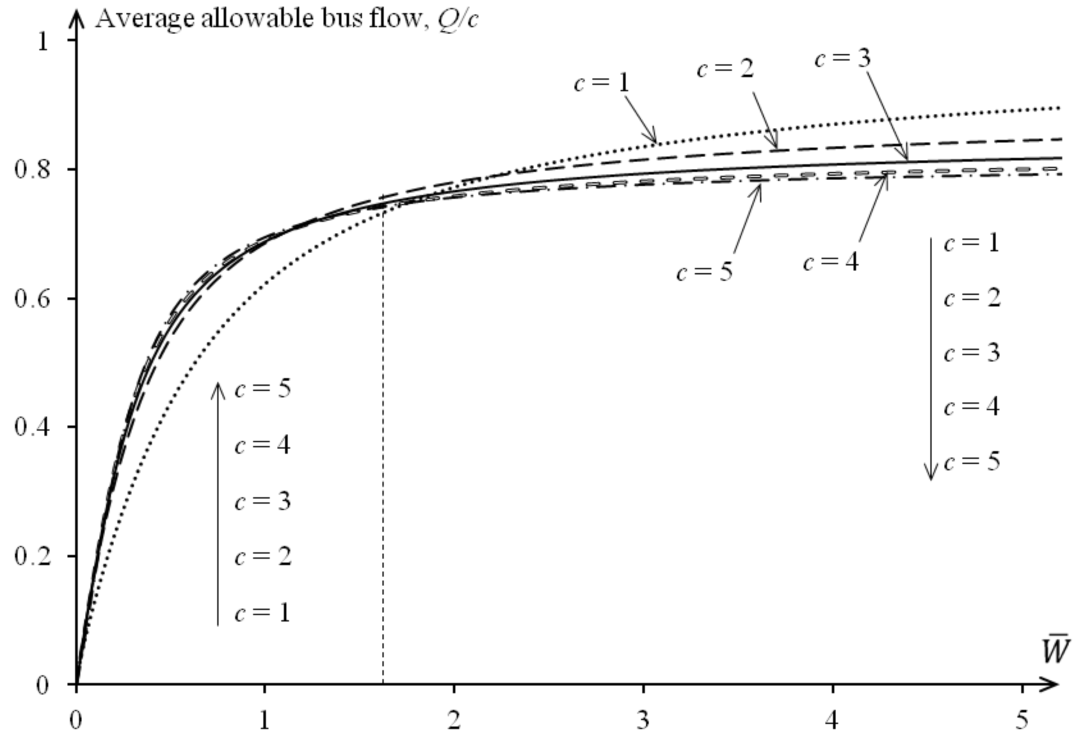
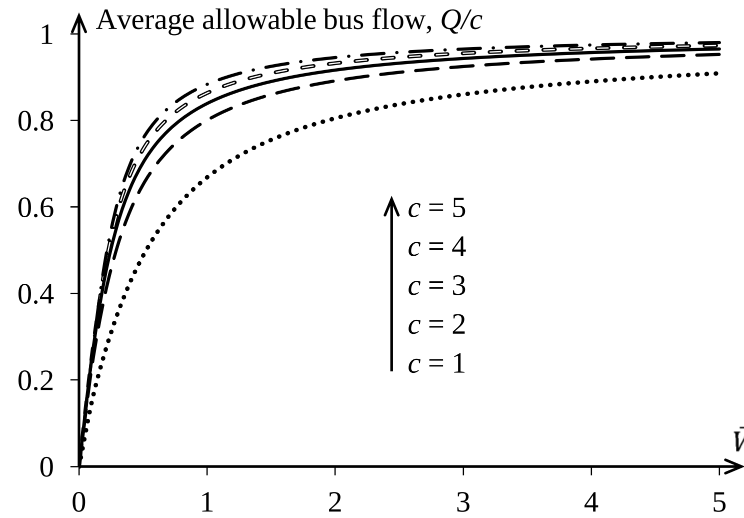


Fig. 3.10 Average allowable bus flow per berth versus $\overline{W}$ ($C_S = 0.2$)

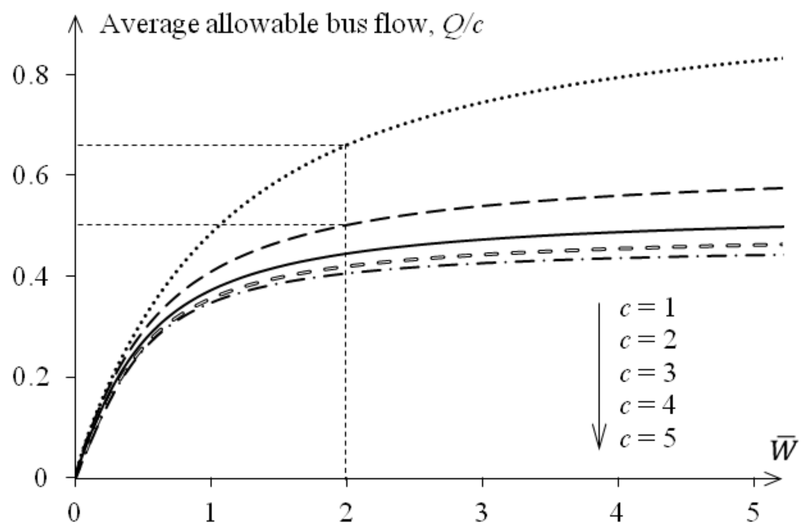
Graphs like Figs. 3.10 or 3.11 can be used in a number of practical pursuits. For example, the curves in these figures can be used to determine the number of berths needed to achieve targets for $\overline{W}$ and Q. Or, they can be used to estimate Q given bus arrival rate and a specified number of berths. The figures can also help determine when it can be advantageous to split a single stop with many berths into multiple adjacent stops. For example, we can see from Fig. 3.11(b) that, when $\overline{W} = 2$, splitting a 2-berth stop into two single-berth stops could increase the total allowable bus flow by 30%. Admittedly, this prediction assumes certain idealizations; e.g., that both the bus arrival processes and the service time distributions are comparable across the stops; and that buses bound for one of these stops do not impede buses bound for the other.

3.5 Effects of Variations in Bus Service Times and Headways

Having explored the influences of $\overline{W}$, we now examine how the returns in Q are influenced by the coefficients of variation in bus service time, C_S, and in bus headway, C_H.



(a) $C_S = 0$



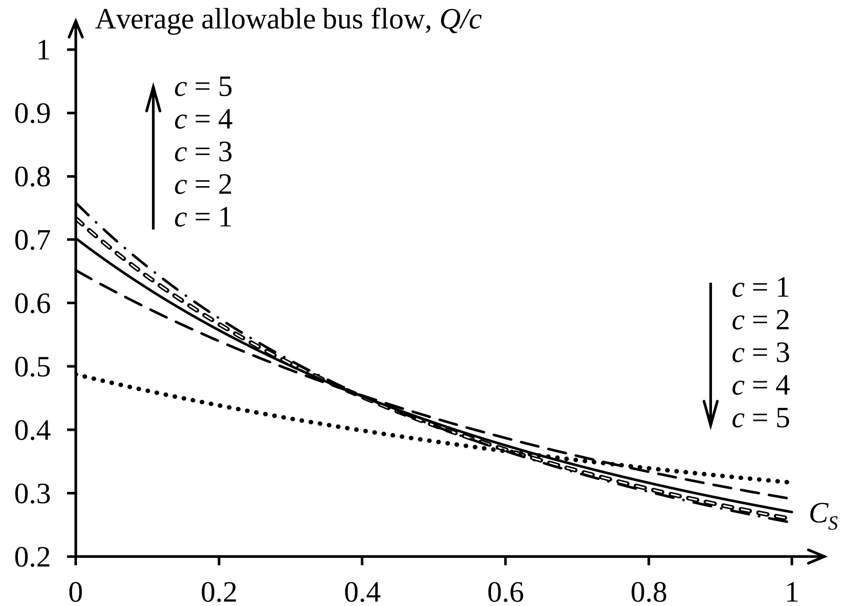
(b) $C_S = 1$

Fig. 3.11 Average allowable bus flow per berth versus $\bar{W}$ ($C_S = 0$ and 1)

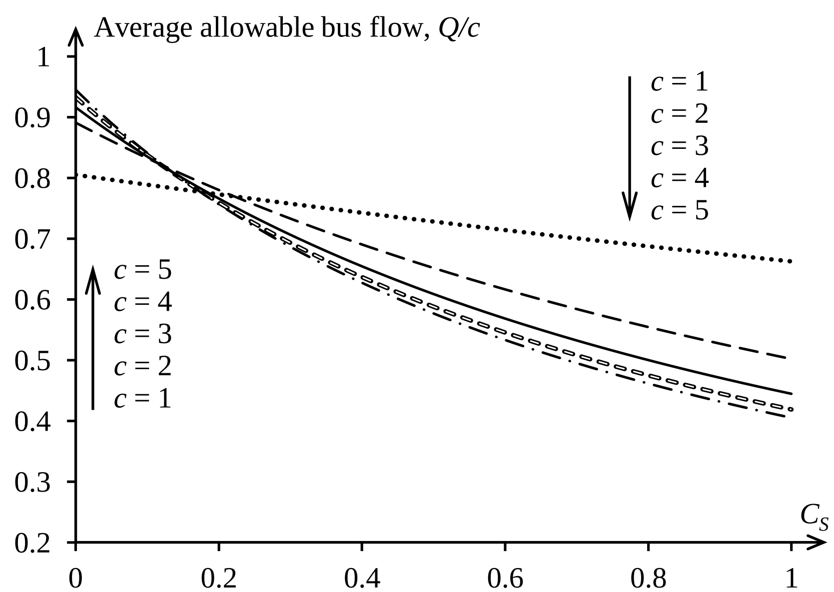
3.5.1 Effects of C_S

We continue to assume that bus arrivals are Poisson and use (3.9) to explore the allowable bus flows for the range of $C_S \in [0, 1]$.

Fig. 3.12(a) displays effects of C_S on the average allowable bus flow for $c = 1$-5 when $\bar{W} = 0.5$. Note from the figure the downward sloping curves (shown with different line-types for distinction). These curves reveal how C_S exerts an inverse influence on Q. Further note that the slopes of the curves increase from $c = 1$ to $c = 5$. This reveals an inverse influence of C_S on the returns in Q. As a result, the stops exhibit increasing returns in Q for low C_S values (thanks to the berth pooling effect) and decreasing returns in Q for high C_S (thanks to the blocking effect). These inverse influences of C_S become more dramatic as $\bar{W}$ increases. To illustrate, the above analysis is repeated, but for $\bar{W} = 2$. Results are displayed in Fig. 3.12(b).



(a) $\bar{W} = 0.5$



(b) $\bar{W} = 2$

Fig. 3.12 Average allowable bus flow per berth versus C_S ($\bar{W} = 0.5$ and 2)

3.5.2 Effects of C_H

Equation (3.9) cannot be used to calculate allowable bus flows for non-Poisson bus arrivals. Hence we use simulation to explore how variations in bus arrival headway affect things. We will assume that bus service time and headway are both gamma-distributed with $C_S = 0.6$ and C_H ranging from 0 to 1. The simulation algorithm is furnished in Appendix C.2.

Fig. 3.13(a) shows effects of C_H on the average allowable bus flow for $c =$ 1-5 when $\bar{W} = 0.5$. Once again, the downward slopes of the curves in Fig. 3.13(a) mean that the allowable flow diminishes as C_H increases. Further note that the slope decreases as c increases. This means that a larger C_H mitigates the trend of decreasing returns in Q; i.e., C_H has a positive effect on the returns in Q. This too is to be expected, because a larger C_H means more fluctuations in bus arrivals, and thus a more significant pooling effect. However, these effects of C_H, diminish as $\bar{W}$ increases; see Fig. 3.12(b) for the case of $\bar{W} = 2$.

3.6 Summary of Findings

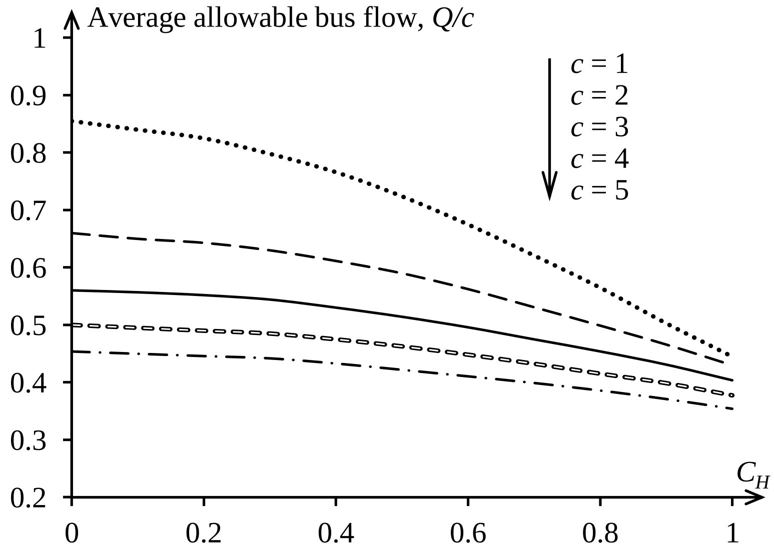
This chapter explored how key operating factors influence the allowable bus flows of no-overtaking stops. These key factors include: the target service level specified by the transit agency, the number of berths, the bus service time distribution and the bus arrival process. Analytical and simulation models were developed to this end. The models account for the influences of these factors in ways that are more complete than what has been offered by formulas in well-known handbooks and other academic literature. Through this more complete accounting come insights. The insights have practical implications.

The contributions of the work presented in this chapter are summarized below:

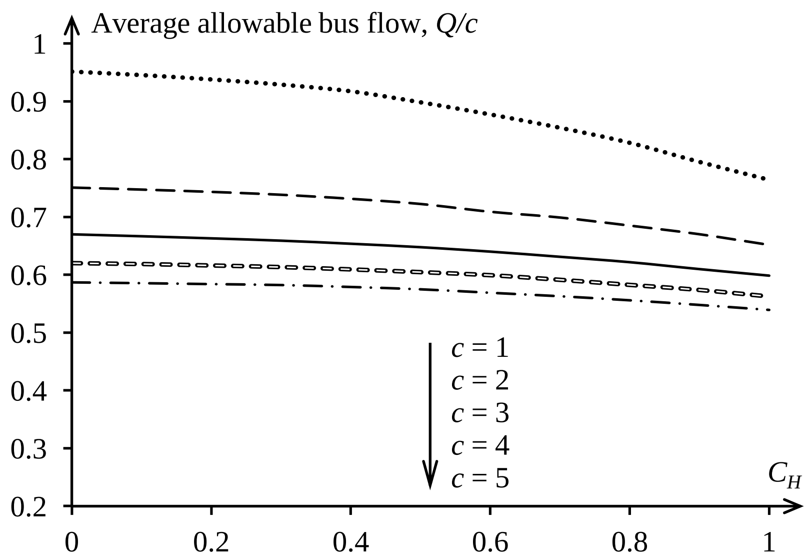
i) Analytical solutions are developed for a queueing model that describes the unique operating features of serial bus berths in a no-overtaking stop. Note that the solution methodology can be applied to other serial queueing systems, e.g., taxi queues, toll stations with tandem booths, and Personal Rapid Transit systems. Further note that no literature on queueing theory has been found to address a model of this kind. From the analytical solutions, a closed-form and parsimonious approximation is derived for predicting the maximum bus flows that can depart a class of bus stops while maintaining target service levels.

ii) Results from the models are largely at odds with those in the literature, including the well-known *HCM*. For example, the models of multi-berth stops predict that adding berths to a stop can sometimes return disproportionally high gains in allowable bus flow. Moreover, the results show that the widely-used service level metric for bus stops, failure rate, is not a good proxy for the delays that buses experience at the stops.

iii) The models can be used to guide transit agencies to determine the suitable number of berths when designing a new stop, and when expanding or splitting an existing, congested bus-stop. The models also unveil possibilities for improving the bus operations at stops by using appropriate operating strategies. These strategies may include means to reduce the variations in bus service time and in bus headway (e.g., technologies for facilitating passenger loading processes at the stop).



(a) $\bar{W} = 0.5$



(b) $\bar{W} = 2$

Fig. 3.13 Average allowable bus flow per berth versus C_H ($\bar{W} = 0.5$ and 2)

The insights in this chapter obtained for the general case confirm the findings from the extreme case, as previously presented in Chapter 2. For example, the allowable bus flow is found to diminish as C_S increases. Further note that the solution methodology can also be applied to limited-overtaking stops with some modifications. (Admittedly, however, the solution of limited-overtaking stops would be more complicated.) And the insights obtained from no-overtaking stops also hold to a greater or lesser degree for limited-overtaking stops. The reader can use the simulation algorithm furnished in Appendix C.3 to verify this.

Chapter 4

Conclusions

Congested bus stops are often the bottlenecks in a bus system. To identify means to mitigate or eliminate the bus queueing at these stops, one needs first to understand how bus queueing is affected by key operating factors, including the stop's number of berths, the bus arrival process and dwell time distribution, and queue disciplines that describe how buses enter/exit a berth. This dissertation has unveiled cause-and-effect relations between these operating features and metrics, such as the average bus delay at the stop. This was done by developing analytical and simulation models that describe the unique operating features of serial bus berths in an isolated stop. Unlike previous works on the subject which are either case-specific or furnish incorrect or incomplete analysis, the models developed in this work provide exact or approximate solutions for a wide range of operating conditions, and thus can be applied to virtually any isolated bus stop.

The contributions of the dissertation are summarized in Section 4.1. Directions for future work are discussed in Section 4.2.

4.1 Contributions

Primary contributions of this dissertation are summarized as follows:

i) Analytical models are developed both, for bus stops operating in the extreme case when bus queues are always present, and in the more general case. These models furnish exact solutions of the average bus delays and the allowable bus flows from the stops. These analytical works also contribute to queueing theory, since no analytical model has previously been found for the unique queueing system that is a bus stop.

ii) Many new insights are obtained from these models. These insights either correct or complete previous findings and conventional wisdom on the bus queueing problem at stops. Important insights include: a) failure rate, the service level metric that is widely used in literature, is not a good proxy for average bus delay, the metric that is more aligned with the concerns of transit users and agencies; b) the allowable bus flow decreases as the variations in bus headway and in bus dwell time increase; c) the returns in allowable bus flow increase with berth number when the target average bus delay is small, and decrease otherwise, and are affected negatively by the variation in bus dwell time, and positively by the variation in bus headway; and d) allowing bus overtaking maneuvers can improve the allowable bus flow when the variation in dwell time is high, but will diminish this flow otherwise.

iii) The models and insights can be put to practical use. They can help practitioners to determine the bus delays at a congested stop, or inversely, to determine the allowable bus flow under a target service level, or the number of berths required to serve a given bus flow. Practitioners can also use the models to choose suitable strategies to improve the bus operations at busy stops. These strategies include: permitting or prohibiting bus overtaking, and means to reduce variations in bus dwell time and in bus headway.

We understand that this dissertation focuses on service level measures for buses only (FR, $\overline{W}$), while people care more about the service level of bus users. For example, passengers waiting at a stop might prefer more buses arriving at the stop, disregarding their queueing delays. Yet our chosen service level metric, the average bus delay, is aligned with the average delay per onboard passenger; and the number of passengers onboard is usually much larger than that waiting at a stop. Nevertheless, passengers waiting at the stop can still benefit from a higher allowable bus flow given a target $\overline{W}$.

To be sure, all of our present models are idealized, particularly since they apply to isolated stops operating in steady state. Yet these models represent a step toward better understanding bus-stop operation. With small modifications, the present models could account for the lost times due to bus deceleration and acceleration into and out of a stop, and the extra times required for boarding passengers as they walk to appropriate berths at a multi-berth stop. The models can also be used to predict other outputs of interest, such as the distribution of bus queue lengths at a stop. Predictions of this latter kind could be especially useful at stops with limited space for storing queued buses. Moreover, these models can be applied to other serial queueing systems, e.g., taxi queues, Personal Rapid Transit stations, and highway toll stations with tandem service berths. Other directions that can build upon the present work are explored next.

4.2 Future Work

A long list of research opportunities can build upon the present work. Some of these opportunities are summarized below:

i) Assumptions can be relaxed for developing better, more realistic analytical models for bus stops. In addition to those discussed at the end of the previous section, we can also relax the assumptions regarding a) the bus arrival process, which was assumed to be Poisson in the present work; and b) the independence between bus headways and dwell times. Our recent work indicates that, at least under certain operating conditions, embedded Markov chains exist when either of these two assumptions is relaxed. Analytical models can therefore be developed.

Work of this kind could yield more realistic bus stop models, since bus arrivals in the real world are not exactly Poisson, and bus dwell times are often correlated with bus headway (Daganzo, 2009).

ii) Strategies can be examined that can improve bus operations at busy stops. These strategies include bus platooning (Szász et al., 1978; Gardner et al., 1991), in which buses arrive and depart stops in batched fashion, and stop splitting, in which a stop with multiple berths is partitioned into multiple neighboring stops. Our recent work shows that bus platooning has the effect of reducing the variation in bus headway, and thus can increase the allowable bus flow at a stop. Stop splitting has been briefly discussed in Chapter 3, but a more complete analysis on this strategy should also consider the changes in bus arrival process and dwell time distribution after splitting the stop. Work of this kind could unveil guidelines on how to optimally apply these strategies under various circumstances.

iii) The work can be extended to stops that are affected by nearby traffic signals. This would also yield more realistic bus stop models because bus stops are often placed short distances from signalized intersections to improve accessibility; e.g., by enabling bus-users to readily transfer between different bus lines (TRB, 1996). The literature on stops of this type either focus on deterministic models (Furth and SanClemente, 2006), or rely on simulation (Gibson, 1996; Kim and Rilett, 2005; Zhou and Gan, 2005). However, analytical models that describe stochastic bus operations at these stops can be developed, at least for the extreme case when a bus queue is always present. We are particularly interested in the results that would be obtained from these proposed models, e.g., how the distance between the stop and the traffic signal affects the stop's allowable bus flow.

Bibliography

Cox, D.R., 1955. The analysis of non-Markov stochastic processes by the inclusion of supplementary variables. *Proceedings of the Cambridge Philosophical Society (Mathematics and Physical Sciences)* 51, 433-441.

Crommelin, C.D., 1932. Delay probability formulae when the holding times are constant. *The Post Office Electrical Engineers' Journal* 25, 41-50.

Daganzo, C.F., 2009. A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. *Transportation Research Part B* 43(10), 913-921.

Danas, A., 1980. Arrivals of passengers and buses at two London bus stops. *Traffic Engineering and Control* 21(10), 472-475.

Erlang, A.K., 1909. The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik B* 20, 33-39.

Estrada, M., Ortigosa, J., Robusté, F., 2011. Tandem bus stop capacity. *TRB 90th Annual Meeting Compendium of Papers* 11-1747. Transportation Research Board of the National Academies, Washington D.C.

Fernandez, R., 2001. A new approach to bus stop modeling. *Traffic Engineering and Control* 42(7), 240-246.

Fernandez, R., 2010. Modelling public transport stops by microscopic simulation. *Transportation Research Part C: Emerging Technologies* 18(6), 856-868.

Fernandez, R., Planzer, R., 2002. On the capacity of bus transit systems. *Transport Reviews* 22(3), 267-293.

Furth, P.G., SanClemente, J.L., 2006. Near side, far side, uphill, downhill: impact of bus stop location on bus delay. *Transportation Research Record: Journal of Transportation Research Board* 1971, Transportation Research Board of the National Academies, Washington D.C., 66-73.

Gardner, G., Cornwell, P.R., Cracknell, J.A., 1991. The Performance of Busway Transit in Developing Cities. *TRRL Research Report* 329, Transport and Road Research Laboratory, Crowthorne.

Ge, H., 2006. Traffic impacts of bus stops in urban area and related optimization techniques. *Ph.D. Thesis*, Southeast University, China. (in Chinese)

Gibson, J., 1996. Effects of a downstream signalized junction on the capacity of a multiple berth bus-stop. In: *Proceedings of the 24th PTRC European Transport Forum*, London.

Gibson, J., Baeza, I., Willumsen, L.G., 1989. Bus stops, congestion, and congested bus stops. *Traffic Engineering and Control* 30(6), 291-302.

Gross, D., Shortle, J.F., Thompson, J.M., Harris, C.M., 2008. *Fundamentals of Queueing Theory*, 4th Ed. John Wiley & Sons, Inc., Hoboken, New Jersey.

Gu, W., Li, Y., Cassidy, M.J., 2012. On the capacity of highway checkpoints: Models for unconventional configurations. *Submitted for journal publication*.

Hall, R.W., Daganzo, C.F., 1983. Tandem toll booths for the Golden Gate Bridge. *Transportation Research Record: Journal of the Transportation Research Board* 905, Transportation Research Board of the National Academies, Washington, D.C., 7-14.

ITE, 1999. *Transportation Planning Handbook*. Institute of Transportation Engineers, New Jersey.

Kendall, D.G., 1953. Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded Markov chains. *The Annals of Mathematical Statistics* 24, 338-354.

Khintchine, A.Y., 1932. Mathematical theory of a stationary queue. *Matematicheskii Sbornik* 39(4), 73-84.

Kim, W., Rilett, L.R., 2005. Improved transit signal priority system for networks with nearside bus stops. *Transportation Research Record: Journal of Transportation Research Board* 1925, Transportation Research Board of the National Academies, Washington D.C., 205-214.

Kohler, U., 1991. Capacity of transit lanes. *Proceedings of the International Symposium on Highway Capacity*, Karlsruhe, Germany.

Little, J.D.C., 1961. A proof of the queuing formula: $L=\lambda W$. *Operations Research* 9(3), 383–387.

Maal øe, E., 1973. Approximation formulae for estimation of waiting-time in multiple-channel queueing system. *Management Science* 19, 703-710.

Neuts, M.F., 1978. Markov chains with applications in queueing theory, which have a matrix-geometric invariant probability vector. *Advances in Applied Probability* 10(1), 185-212.

Newell, G.F., 1982. Applications of Queueing Theory, Second ed. London: Chapman and Hall.

Nozaki, S.A., Ross., S.M., 1975. Approximations in multi-server Poisson queues. *Report*, University of California, Berkeley.

Papacostas, C.S., 1982. Capacity characteristics of downtown bus streets. *Transportation Quarterly* 36(4), 617-630.

Pollaczek, F., 1930. Über eine Aufgabe der Wahrscheinlichkeitstheorie. *Mathematische Zeitschrift* 32, 64-100.

St. Jacques, K.R., Levinson, H.S., 1997. Operational Analysis of Bus Lanes on Arterials. *TCRP Report* 26, Transportation Research Board of the National Academies, Washington D.C.

Szász, P.A., Montana, L.C., Ferreira, E.O., 1978. COMONOR: ordinated bus convoy. *Technical Paper* 9, Companhia de Engenharia de Trafego, Sao Paulo.

TRB, 1985. *Highway Capacity Manual*. Transportation Research Board of the National Academies, Washington D.C.

TRB, 1996. Guidelines for the location and design of bus stops. *TCRP Report* 19, Transportation Research Board of the National Academies, Washington D.C.

TRB, 2000. *Highway Capacity Manual*. Transportation Research Board of the National Academies, Washington D.C.

TRB, 2003. *Transit Capacity and Quality of Service Manual*, 2nd Ed. Transportation Research Board of the National Academies, Washington D.C.

Whitt, W., 1993. Approximations for the GI/G/m queue. *Production and Operations Management* 2(2), 114-161.

Wolff, R.W., 1982. Poisson arrivals see time averages. *Operations Research* 30(2), 223-231.

Zhou, G., Gan, A., 2005. Performance of transit signal priority with queue jumper lanes. *Transportation Research Record: Journal of Transportation Research Board* 1925, Transportation Research Board of the National Academies, Washington D.C., 265-271.

Appendix A

Glossary of Symbols

$\bar{A}$ – average of A_n

A_n – number of bus arrivals in cycle n in the general case

C_S – coefficient of variation in bus service time

$C_{S,crit}$ – critical coefficient of variation in bus service time, where the allowable flows for limited-overtaking and no-overtaking stops are the same

C_H – coefficient of variation in bus headway

c – number of berths in a stop

$\mathcal{D}$ – queue discipline at a stop

FR – failure rate

$F_S(t)$ – cumulative distribution function of bus service time

$\mathcal{L}$ – service level metric (e.g., failure rate or average bus delay) for a stop

$\tilde{L}_n$ – number of buses queued at the stop's entrance at the n-th regenerative point in the general case

$\bar{L}_q$ – average number of buses in queue over time

M_n – number served during cycle n in the general case

$\bar{N}$ – average number of buses served in a cycle defined in Section 2.2.1

$\bar{N}_{SC}$ – average number of subcycles in a cycle; subcycle and cycle are defined in Section 2.2.2

$\bar{n}$ – average number of buses served per cycle, save for those cycles with specified n_0 and t_E

n_0 – number of buses that have been served or are under service in the present cycle at a control point

n_1 – expected number of buses that would be served in the cycle if one decides to block any queued buses from entering empty upstream berth(s) at a control point

n_2 – expected number of buses that would be served in the cycle if one decides to allow queued buses to enter empty upstream berth(s) at a control point

$\mathcal{P}_H$ – distribution of bus headway

$\mathcal{P}_S$ – distribution of bus service time

$\boldsymbol{P}$ – transition matrix of the Markov chain for the general case

$P_{i,j}$ - transition probability of the Markov chain for the general case

p – probability that a cycle with specified n_0 and t_E occurs

Q – bus flow, allowable bus flow

Q_{extr} – allowable bus flow in the extreme case when $\overline{W} = \infty$ and $FR = 1$

Q_{LOT_extr} – allowable bus flow for a limited-overtaking stop in the extreme case

Q_{NOT_extr} – allowable bus flow for a no-overtaking stop in the extreme case

q - average bus discharge flow of all the cycles except for those with specified n_0 and t_E

S - bus service time

$\bar{S}$ – mean of bus service time

S' – service time of the bus that is allowed to be served at the upstream berth of a 2-berth stop at a control point

$\bar{T}$ – average duration of a cycle defined in Section 2.2.1

T_{SC} – duration of a subcycle defined in Section 2.2.2

$\overline{TL}$ – average of TL_n

TL_n – sum of the queue lengths seen by each bus arrival in cycle n in the general case

$\bar{t}$ – average cycle length, save for those cycles with specified n_0 and t_E

t_E – elapsed service time of the bus dwelling at the downstream-most berth at a control point

t_{E_crit} – critical elapsed service time of the bus dwelling at the downstream-most berth at a control point; the decision is to block the bus if $t_E > t_{E\_crit}$

t_R – remaining service time of the bus dwelling at the downstream-most berth at a control point

t_1 – expected cycle length if one decides to block any queued buses from entering empty upstream berth(s) at a control point

t_2 – expected cycle length if one decides to allow queued buses to enter empty upstream berth(s) at a control point

$\bar{W}$ – average bus delay

$\bar{W}_b$ – average bus delay in the berth after the bus has served its passengers

$\bar{W}_q$ – average bus delay in the entry queue

$\bar{w}$ – un-normalized target average bus delay

Δn – difference between the extra bus numbers served in a cycle between the two control decisions made at a control point (blocking the bus and letting the bus in)

λ – bus arrival rate

ρ – ratio of bus inflow to the supremum of the bus discharge flow from the stop

$\boldsymbol{\pi}$ – limiting probability distribution of the Markov chain for the general case

π_i – limiting probability of the Markov chain for the general case

$\tilde{\pi}(z)$ – z-transform of $\boldsymbol{\pi}$

Appendix B

Mathematical Proofs and Derivations

B.1 Proof of Equation (2.2)

For a no-overtaking stop with c berths, let $T(c) = max_j\{S_j\}, (j = 1,2,\ldots,c)$ be the service time of a bus platoon (recall that a bus queue is always present upstream of the stop), where S_j is the service time of the j-th bus in the platoon. All S_j's are i.i.d. random variables subject to the CDF $F_S(t)$. Let $F_{T(c)}(t)$ be the CDF of $T(c)$, we have:

$$F_{T(c)}(t) = P\{T(c) \leq t\} = P\{S_j \leq t, j = 1,2,\ldots,c\}$$

$$= \prod_{j=1}^{c} P\{S_j \leq t\}$$

$$= \prod_{j=1}^{c} F_S(t) = (F_S(t))^c \, .$$

From the identity $E[T(c)] = \int_{t=0}^{\infty}\left(1 - F_{T(c)}(t)\right)dt$, we have:

$E[T(c)] = \int_{t=0}^{\infty}(1 - (F_S(t))^c)dt$. Thus,

$$Q_{NOT\_extr} = \frac{c}{E[T(c)]} = \frac{c}{\int_{t=0}^{\infty}(1-(F_S(t))^c)dt} \, . \qquad \blacksquare$$

B.2 Proof of Equation (2.6) for Erlang-Distributed Bus Service Times

We instead prove a general result: if random variables A and B follow Erlang distributions with parameters (λ, m) and (λ, n), respectively, then

$$Pr\{A > B\} = 2^{1-m-n} \cdot \sum_{j=n}^{m+n-1} \binom{m+n-1}{j} \, .$$

From the definition of Erlang distribution, we can let $A = \sum_{j=1}^{m} A_j$, $B = \sum_{j=1}^{n} B_j$, where A_j's and B_j's are independent, exponentially distributed random variables with parameter λ.

Consider a boxing game between two teams; team $\mathscr{A}$ has m players and team $\mathscr{B}$ has n. Each team dispatches its players one by one to the arena. At the end of a round, only the winner stays in the arena for the next round. The one who is knocked out will not be allowed to enter the game again (thus his team has to dispatch a new player, if any left). A team wins the game when the other team is wiped out.

Now consider that A_j and B_j represent the initial strengths of the j-th players of the two teams. If team $\mathscr{A}$'s j-th (fresh) player fights against team $\mathscr{B}$'s i-th (fresh) player, the former wins if and only if $A_j > B_i$, and the remaining strength of the winner at the end of the round will be $A_j - B_i$. So team $\mathscr{A}$ wins if and only if $A > B$.

Next we calculate the probability that team $\mathscr{A}$ wins the game in a different way. First note that due to the memoryless property of the exponential distribution, a player who wins a round will remain like new in the next round. Thus in any round, a player wins with a probability of 0.5. Further note that "team $\mathscr{A}$ wins" means that they can lose at most $m-1$ players before they wipe out the n players of team $\mathscr{B}$. Since the number of team $\mathscr{A}$'s losses before it achieves n wins follows the negative binomial distribution, we have

$$Pr\{A > B\} = Pr\{\text{team } \mathscr{A} \text{ wins}\} = \sum_{j=n}^{m+n-1} \binom{m+n-1}{j} / 2^{m+n-1}.$$

Now if S_j is an Erlang-k random variable, $\sum_{j=2}^{i+1} S_j$ will be an Erlang-ik random variable with the same parameter λ. Hence,

$$Pr\left\{S_1 > \sum_{j=2}^{i+1} S_j\right\} = 2^{1-(i+1)k} \cdot \sum_{j=ik}^{(i+1)k-1} \binom{(i+1)k-1}{j}. \qquad \blacksquare$$

B.3 Proof of Equation (2.8)

We only need show:

$$\sum_{i=1}^{\infty} 2^{1-(i+1)k} \cdot \sum_{j=ik}^{(i+1)k-1} \binom{(i+1)k-1}{j} = \sum_{j=1}^{k} \frac{2\cdot(-1)^{k-j}}{(k-j)!} \cdot \left. \frac{d^{k-j}\left(\frac{x^{-(j+1)}}{x^k-1}\right)}{dx^{k-j}} \right|_{x=2}. \qquad \text{(B.3.1)}$$

By interchanging the two summation operators on the left hand side, we have:

$$\sum_{i=1}^{\infty} 2^{1-(i+1)k} \cdot \sum_{j=ik}^{(i+1)k-1} \binom{(i+1)k-1}{j} = \sum_{j=1}^{k} \sum_{i=1}^{\infty} 2^{1-(i+1)k} \cdot \binom{(i+1)k-1}{ik+j-1}$$

$$= \sum_{j=1}^{k} \sum_{i=1}^{\infty} 2^{1-(i+1)k} \cdot \frac{\left((i+1)k-1\right)!}{(ik+j-1)! \cdot (k-j)!}$$

$$= \sum_{j=1}^{k} G_j(2), \qquad \text{(B.3.2)}$$

where $G_j(x) = \sum_{i=1}^{\infty} x^{1-(i+1)k} \cdot \frac{((i+1)k-1)!}{(ik+j-1)! \cdot (k-j)!}$.

We define

$$F_j(x) = \sum_{i=1}^{\infty} x^{-(ik+j)} = \frac{x^{-(j+1)}}{x^k - 1}.$$

The last equality holds when $|x| > 1$. We have

$$\frac{d^{k-j} F_j(x)}{dx^{k-j}} = \sum_{i=1}^{\infty} (-1)^{k-j} \cdot \left[(ik+j) \cdot (ik+j+1) \cdot \ldots \cdot \left((i+1)k-1\right)\right] \cdot x^{-(i+1)k}.$$

Thus,

$$G_j(x) = \frac{(-1)^{k-j} x}{(k-j)!} \cdot \frac{d^{k-j} F_j(x)}{dx^{k-j}}. \tag{B.3.3}$$

Combining (B.3.2) and (B.3.3), we have (B.3.1). ∎

B.4 Probability Distribution of T_{SC} (Section 2.2.2)

We first develop the general formula for the CDF, $F_{SC}(t)$, and the probability density function (PDF), $f_{SC}(t)$, of T_{SC}; then we obtain a simplified form for the case of Erlang-distributed bus service times.

$$F_{SC}(t) = Pr\,\{T_{SC} \le t\}$$

$$= \sum_{n=1}^{\infty} Pr\,\{T_{SC} \le t, N = n+1\}$$

(where N denotes the number of buses served in the subcycle)

$$= \sum_{n=1}^{\infty} Pr\,\{\sum_{j=2}^{n} S_j < S_1 \le \sum_{j=2}^{n+1} S_j \le t\}$$

(where S_j $(j = 1,2,\ldots)$ denotes the service time of the j-th bus that enters the subcycle)

$$= \sum_{n=1}^{\infty} \left(Pr\left\{\sum_{j=2}^{n} S_j < S_1 \le t, \sum_{j=2}^{n+1} S_j \le t\right\} - Pr\{\sum_{j=2}^{n+1} S_j < S_1 \le t\}\right)$$

$$= \sum_{n=1}^{\infty} \left(E\left[Pr\left\{\sum_{j=2}^{n} S_j < S_1 \le t, \sum_{j=2}^{n+1} S_j \le t \middle| S_1, S_2, \ldots, S_n, \sum_{j=2}^{n} S_j < S_1 \le t\right\}\right] - Pr\{\sum_{j=2}^{n+1} S_j < S_1 \le t\}\right)$$

$$= \sum_{n=1}^{\infty} \left(\int_{\substack{S_1, S_2, \ldots, S_n, \\ \sum_{j=2}^{n} S_j < S_1 \le t}} Pr\left\{S_{n+1} \le t - \sum_{j=2}^{n} S_j\right\} \cdot dF_n(S_1, S_2, \ldots, S_n) - Pr\{\sum_{j=2}^{n+1} S_j < S_1 \le t\} \right)$$

(where $F_n(S_1, S_2, \dots, S_n)$ denotes the joint CDF of $S_1, S_2, \dots, S_n$)

$$= \sum_{n=1}^{\infty}\left(\int_{\substack{S_1,S_2,\dots,S_n,\\ \sum_{j=2}^{n} S_j < S_1 \le t}} F_S(t - \sum_{j=2}^{n} S_j) \cdot dF_n(S_1, S_2, \dots, S_n) - \int_{R_n=0}^{t} (F_S(t) - F_S(R_n)) \cdot d\Phi_n(R_n)\right)$$

(where $\Phi_n(R_n)$ denotes the CDF of $R_n = \sum_{j=2}^{n+1} S_j$)

$$= F_S(t)^2 + \sum_{n=2}^{\infty}\left(\int_{R_{n-1}=0}^{t} F_S(t - R_{n-1}) \cdot (F_S(t) - F_S(R_{n-1})) \cdot d\Phi_{n-1}(R_{n-1})\right) - \sum_{n=1}^{\infty}\left(\int_{R_n=0}^{t} (F_S(t) - F_S(R_n)) \cdot d\Phi_n(R_n)\right)$$

(where $\Phi_{n-1}(R_{n-1})$ denotes the CDF of $R_{n-1} = \sum_{j=2}^{n} S_j$; and note that $\sum_{j=2}^{n} S_j = 0$ when $n = 1$)

$$= F_S(t)^2 + \int_{s=0}^{t} \left(1 - F_S(t-s)\right) \cdot \left(F_S(t) - F_S(s)\right) \cdot \left(\sum_{n=1}^{\infty} \varphi_n(s)\right) \cdot ds \qquad \text{(B.4.1)}$$

(where $\varphi_n(s)$ denotes the PDF of $s = \sum_{j=1}^{n} S_j$.)

By applying Leibniz integral rule, we have:

$$f_{SC}(t) = 2F_S(t) f_S(t) + \int_{s=0}^{t} \left[-f_S(t-s) \cdot \left(F_S(t) - F_S(s)\right) + \left(1 - F_S(t-s)\right) \cdot f_S(t)\right] \cdot \left(\sum_{n=1}^{\infty} \varphi_n(s)\right) \cdot ds\ , \qquad \text{(B.4.2)}$$

where $f_S(t)$ is the PDF of bus service time.

Now suppose that S_j follows Erlang-k distribution ($k \ge 1$), then $\varphi_n(s)$ is the PDF of the Erlang-nk distribution:

$$\varphi_n(s) = \frac{k^{nk} s^{nk-1} e^{-ks}}{(nk-1)!}\ ;$$

note that the mean service time is 1. Thus,

$$\sum_{n=1}^{\infty} \varphi_n(s) = k e^{-ks} \cdot \sum_{n=1}^{\infty} \frac{(ks)^{nk-1}}{(nk-1)!}\ . \qquad \text{(B.4.3)}$$

We will now use the following equality without proof (the reader can prove it using Taylor's expansion):

$$\frac{1}{k} \sum_{j=1}^{k} e^{x e^{i\frac{2\pi j}{k}}} = 1 + \frac{x^k}{k!} + \frac{x^{2k}}{(2k)!} + \cdots \ \text{(for } k \ge 1\text{)}, \qquad \text{(B.4.4)}$$

where i is the imaginary unit.

We differentiate both sides of (B.4.4) with respect to x:

$$\frac{1}{k}\sum_{j=1}^{k} e^{i\frac{2\pi j}{k}+xe^{i\frac{2\pi j}{k}}} = \frac{x^{k-1}}{(k-1)!} + \frac{x^{2k-1}}{(2k-1)!} + \cdots . \tag{B.4.5}$$

Comparing the right hand side of (B.4.3) and (B.4.5), we have:

$$\sum_{n=1}^{\infty} \varphi_n(s) = ke^{-ks} \cdot \left(\frac{1}{k}\sum_{j=1}^{k} e^{i\frac{2\pi j}{k}+kse^{i\frac{2\pi j}{k}}}\right)$$

$$= \sum_{j=1}^{k} e^{i\frac{2\pi j}{k}+ks\left(e^{i\frac{2\pi j}{k}}-1\right)} . \tag{B.4.6}$$

Equation (B.4.6) can be plugged in (B.4.1) and (B.4.2) to determine the distribution functions numerically. ■

B.5 Proof of Equation (2.16)

From (2.11-15), we have:

$$\Delta n = q(t_2 - t_1) - 1 \geq 0 ,$$

where

$$t_1 = E[S_1 | S_1 \geq t_E] , \text{ and}$$

$$t_2 = E[max\{S_1, t_E + S'\} | S_1 \geq t_E] .$$

Thus,

$$\Delta n = q \cdot (E[max\{S_1, t_E + S'\} | S_1 \geq t_E] - E[S_1 | S_1 \geq t_E]) - 1$$

$$= q \cdot E[max\{0, t_E + S' - S_1\} | S_1 \geq t_E] - 1$$

$$= q \cdot \frac{\int_{t=t_E}^{\infty} f_S(t)dt \cdot \int_{s=t-t_E}^{\infty} f_S(s)(s-t+t_E)ds}{1-F_S(t_E)} - 1 ,$$

where $f_S(t)$ is the PDF of bus service time.

Now consider

$$\int_{s=t-t_E}^{\infty} f_S(s)(s-t+t_E)ds = -\int_{s=t-t_E}^{\infty} (s-t+t_E)d\big(1-F_S(s)\big)$$

$$= -(s - t + t_E)\big(1 - F_S(s)\big)\Big|_{s=t-t_E}^{\infty} + \int_{s=t-t_E}^{\infty}\big(1 - F_S(s)\big)ds$$

$$= \int_{s=t-t_E}^{\infty}\big(1 - F_S(s)\big)ds \ .$$

Thus,

$$\Delta n = q \cdot \frac{\int_{t=t_E}^{\infty} f_S(t)dt \cdot \int_{s=t-t_E}^{\infty}\big(1-F_S(s)\big)ds}{1-F_S(t_E)} - 1$$

$$= q \cdot \frac{\int_{s=0}^{\infty}\int_{t=t_E}^{t_E+s} f_S(t)dt \cdot \big(1-F_S(s)\big)ds}{1-F_S(t_E)} - 1 \ \text{(interchanging the integrals)}$$

$$= q \cdot \frac{\int_{s=0}^{\infty}\big(1-F_S(s)\big)\big(F_S(t_E+s)-F_S(t_E)\big)ds}{1-F_S(t_E)} - 1$$

$$= q \cdot \frac{\int_{s=0}^{\infty}\big(1-F_S(s)\big)\big(F_S(t_E+s)\big)ds - F_S(t_E)}{1-F_S(t_E)} - 1 \ \left(\text{using } \int_{s=0}^{\infty}\big(1 - F_S(s)\big)ds = \bar{S} = 1\right)$$

Hence we have proved (2.16). ■

B.6 Analytical Solution to a Single-Berth Stop with Uniform Bus Arrivals and Erlang-k Service Time (Section 3.1)

Here we furnish a solution by applying a more general result given by Gross, et al. (2008) for a queueing system with generalized-Erlang distributed headways and service time ($GE_j/GE_k/1$, where GE_j and GE_k are the distributions of bus headway and bus service time, respectively).[15] This general result is:

$$\bar{W}_q(s) = \frac{\prod_{m=1}^{k}\left[\left(-\frac{s_m}{\mu_m}\right)(s+\mu_m)\right]}{s\prod_{m=1}^{k}(s-s_m)}, \tag{B.6.1}$$

where $\bar{W}_q(s)$ is the Laplace-Stieltjes transform of the cumulative distribution function (CDF) of bus waiting time; μ_m $(m = 1, \dots, k)$ is the rate of the m-th exponential component of the GE_k distribution; and s_m $(m = 1, \dots, k)$ is the m-th complex root with negative real parts of the following equation with argument s:

$$\prod_{m=1}^{j}\lambda_m\prod_{m=1}^{k}\mu_m - \prod_{m=1}^{j}(\lambda_m - s)\prod_{m=1}^{k}(\mu_m + s) = 0 \ . \tag{B.6.2}$$

[15] A generalized Erlang distribution is the convolution of independent but not necessarily identical exponential random variables. Here a bus headway can be expressed as the sum of j exponential components that are independent but may not be identical; and a bus service time can be expressed as the sum of k such components.

Equation (B.6.2) is also given in Gross, et al. (2008). The λ_m $(m = 1, \dots, j)$ is the rate of the m-th exponential components of the GE_j distribution.

Since the means of the headway and the service time are $\sum_{m=1}^{j} \frac{1}{\lambda_m}$ and $\sum_{m=1}^{k} \frac{1}{\mu_m}$, respectively, we set $\lambda_m(m = 1, \dots, j) = jQ$, $\mu_m(m = 1, \dots, k) = k$ so that the bus headway and the service time are Erlang-j and Erlang-k distributed, respectively; and so that the bus arrival rate and the service rate are Q and 1, respectively (note that the service rate is the reciprocal of the mean bus service time). Given that when j approaches infinity, the limit of the Erlang-j distribution is a deterministic value. We let $j \to \infty$, so that the headway becomes constant. Then (B.6.2) becomes:

$$\left(\frac{k}{k+s}\right)^k = e^{-\frac{s}{Q}} .$$

The solution of the above equation is:

$$s_m = -kQ \cdot \text{LambertW}\left(-\frac{1}{Q} e^{-\frac{1}{Q}-\frac{2(m-1)\pi}{k}i}\right) - 1, (m = 1, \dots, k) , \tag{B.6.3}$$

where function LambertW$(\cdot)$ is the inverse function of $f(w) = we^w$, which is multi-valued in the field of complex numbers, and has no closed-form expression; and i is the imaginary unit.

By picking up the roots of s_m's with negative real parts, plugging them into (B.6.1), and then taking a partial-fraction expansion, we obtain:

$$\bar{W}_q(s) = \frac{\prod_{m=1}^{k}\left[(-s_m)(\frac{s}{k}+1)\right]}{s\prod_{m=1}^{k}(s-s_m)} = \frac{1}{s} - \sum_{m=1}^{k} \frac{\beta_m}{s-s_m} , \tag{B.6.4}$$

where β_m are constant coefficients to be determined by:

$$\beta_m = \prod_{n=1, n\neq m}^{k} \left(\frac{-s_n}{s_m - s_n}\right) (\frac{s_m}{k} + 1)^k . \tag{B.6.5}$$

By applying the inverse Laplace transform on (B.6.4), we obtain the CDF of the bus waiting time:

$$W_q(t) = 1 - \sum_{m=1}^{k} \beta_m e^{s_m t}.$$

Therefore the failure rate becomes

$$FR = 1 - W_q(0) = \sum_{m=1}^{k} \beta_m ; \tag{B.6.6}$$

and the average bus delay is

$$\bar{W} = \int_{t=0}^{\infty} \left(1 - W_q(t)\right) dt = -\sum_{m=1}^{k} \frac{\beta_m}{s_m} . \tag{B.6.7}$$

For any given $k = {C_S}^{-2}$, the right-hand-sides of (B.6.6) and (B.6.7) are functions of Q. Thus we find the relation between the service level metrics and Q. The results can be obtained numerically.

Note that this solution procedure can be easily adapted to solve an $E_j/E_k/1$ system, as examined in Section 3.1.3. ■

B.7 Derivation of (3.8) for Uniformly Distributed Service Time

Suppose that the bus service time follows a uniform distribution in $[1-\alpha, 1+\alpha]$ and recall that the mean service time is normalized to 1. Then $C_S = \frac{\alpha}{\sqrt{3}}$. The CDF of the service time is

$$F_S(t) = \begin{cases} 0, & \text{if } t < 1-\alpha \\ \frac{t-(1-\alpha)}{2\alpha}, & \text{if } 1-\alpha < t < 1+\alpha \\ 1, & \text{if } t > 1+\alpha \end{cases}.$$

From Equation (2.2), we calculate the supremum of bus discharge rate from the stop:

$$Q(c) = \frac{c}{\int_{t=0}^{\infty}(1-(F_S(t))^c)\mathrm{d}t}$$

$$= \frac{c}{\int_{t=0}^{1-\alpha} dt + \int_{t=1-\alpha}^{1+\alpha}\left(1-(\frac{t-(1-\alpha)}{2\alpha})^c\right)\mathrm{d}t}$$

$$= \frac{c}{1+\alpha\frac{c-1}{c+1}} = \frac{c}{1+\sqrt{3}C_S\frac{c-1}{c+1}}.$$

Therefore,

$$Q = \rho \cdot Q(c) = c\rho/(1+\sqrt{3}C_S\frac{c-1}{c+1}).$$ ■

Appendix C
Simulation Algorithms

C.1 Allowable Bus Flow for a 2-Berth Limited-Overtaking Stop when the General Control Strategy is Applied

The variables used in this simulation are listed as follows:

S_i – Service time of the i-th bus ($i = 1, 2, 3, ...$), not including the time that the bus waits to depart the stop after it has finished serving passengers;

P_i – Position (number) of the berth where the i-th bus dwells to serve passengers; the downstream and the upstream berths are numbered 1 and 2, respectively;

D_i – Time when the i-th bus departs from the stop;

AVL_j – Time when the j-th berth ($j = 1, 2$) becomes available;

T_{start} – Start time of the current bus dwelling at the downstream berth;

M – Maximum number of buses;

I – Intermediate variable.

The algorithm of the simulation model is as follows:

Step 1: For a given bus service time distribution, generate a random variable sequence $\{S_i\}$. Set AVL_j ($j = 1, 2$) and T_{start} to 0. Set $i = 1$.

Step 2: $I = \begin{cases} 1, & \text{if } AVL_1 \leq AVL_2 \\ 2, & \text{otherwise} \end{cases}$.

Step 3: If $I = 1$, or if $AVL_2 - T_{start} > t_{E\_crit}$, go to Step 4; otherwise go to Step 5.

Step 4: $P_i = 1$;

$T_{start} = AVL_2 = AVL_1$;

$AVL_1 = AVL_1 + S_i$;

Go to Step 6.

Step 5: $P_i = 2$;

$AVL_2 = AVL_2 + S_i$;

$AVL_1 = \max\{AVL_1, AVL_2\}$.

Step 6: $D_i = AVL_{P_i}$.

Step 7: If $i = M$, go to Step 8; otherwise, set $i = i + 1$, and go to Step 2.

Step 8: The allowable flow is given by M/D_M. ■

C.2 Allowable Bus Flow for a No-Overtaking Stop in the General Case

New variables used in this simulation are listed as follows (the rest of the notation is the same as in C.1):

H_i – Headway between the arrivals of the $(i-1)$-th bus and the i-th bus; H_1 is the system idle time before the first bus arrives;

$W_{i,q}$ – Waiting time in the queue of the i-th bus before it enters the stop;

$W_{i,b}$ – Waiting time in the berth of the i-th bus after its service is finished;

W_i – Total waiting time of the i-th bus;

F_i – Indicator that takes 1 if the i-th bus fails to enter the berth immediately upon its arrival to the stop, and 0 otherwise.

The algorithm of the simulation model is as follows:

Step 1: For given distributions of bus service time and bus headway, generate random variable sequences $\{S_i\}$ and $\{H_i\}$.

Step 2: Set $i = 1$, $P_i = 1$, $W_{i,q} = W_{i,b} = W_i = F_i = 0$. Set $i = i + 1$.

Step 3: $W_{i,q} = \begin{cases} \max \ \{0, W_{i-1,q} + S_{i-1} + W_{i-1,b} - H_i\}, & \text{if } P_{i-1} = c \\ \max \ \{0, W_{i-1,q} - H_i\}, & \text{otherwise} \end{cases}.$

Step 4: $P_i = \begin{cases} 1, & \text{if } W_{i-1,q} + S_{i-1} + W_{i-1,b} - H_i - W_{i,q} \le 0 \\ P_{i-1} + 1, & \text{otherwise} \end{cases}.$

Step 5: $W_{i,b} = \max \ \{0, W_{i-1,q} + S_{i-1} + W_{i-1,b} - H_i - W_{i,q} - S_i\}$.

Step 6: $W_i = W_{i,q} + W_{i,b}$.

Step 7: $F_i = \begin{cases} 1, & \text{if } W_{i,q} > 0 \\ 0, & \text{if } W_{i,q} = 0 \end{cases}.$

Step 8: If $i = M$, go to Step 9; otherwise, set $i = i + 1$, and go to Step 3.

Step 9: The average bus delay and failure rate are calculated by averaging $\{W_i\}$ and $\{F_i\}$, respectively. The allowable bus flow is given by the inverse of the mean of bus headway. ■

C.3 Allowable Bus Flow for a Limited-Overtaking Stop in the General Case

New variables used in this simulation are listed as follows (the rest of the notation is the same as in C.1-2):

ARR_i – arrival time to the stop of the i-th bus.

The algorithm of the simulation model is as follows:

Step 1: For given distributions of bus service time and bus headway, generate random variable sequences $\{S_i\}$ and $\{H_i\}$.

Step 2: Set $i = 1$, $ARR_i = H_i$, $P_i = 1$, $W_i = F_i = 0$.

Step 3: Set $AVL_1 = H_1 + S_1$; $AVL_j = 0$ for $j = 2, \dots, c$. Set $i = i + 1$.

Step 4: $ARR_i = H_i + ARR_{i-1}$.

Step 5: $W_i = \begin{cases} \max \ \{0, W_{i-1} + S_{i-1} - H_i\}, & \text{if } P_{i-1} = c \\ \max \ \{0, W_{i-1} - H_i\}, & \text{otherwise} \end{cases}.$

Step 6: $P_i = \min \ \{j | AVL_j \le ARR_i + W_i, j = 1, \dots, c\}$.

Step 7: $AVL_{P_i} = ARR_i + W_i + S_i$.

Step 8: Set $AVL_j = \max \{AVL_j, AVL_{P_i}\}$ for $j = 1,2, \ldots, P_i - 1$.

Step 9: $F_i = \begin{cases} 1, \text{if } W_i > 0 \\ 0, \text{if } W_i = 0 \end{cases}$.

Step 10: If $i = M$, go to Step 11; otherwise, set $i = i + 1$, and go to Step 4.

Step 11: The average bus delay and failure rate are calculated by averaging $\{W_i\}$ and $\{F_i\}$, respectively. The allowable bus flow is given by the inverse of the mean of bus headway. ■

Appendix D

Analytical Solution for M/D/c/SRL and M/G/2/SRL Systems

D.1 M/D/c/SRL – Transition Probabilities

We obtain the transition probabilities, $P_{i,j}$, by conditioning on the number of buses that get served in the n-th cycle, M_n, and then by finding all the $Pr\{\tilde{L}_{n+1} = j, M_n = k | \tilde{L}_n = i\}$ for $1 \leq k \leq c$. These $Pr\{\tilde{L}_{n+1} = j, M_n = k | \tilde{L}_n = i\}$ are also useful when calculating the average bus delay (see Section 3.2.4). Note that service times are normalized to 1 for this system.

First, we obtain the following equations by conditioning:

$$P_{0,0} = \sum_{k=1}^{c} Pr\{\tilde{L}_{n+1} = 0, M_n = k | \tilde{L}_n = 0\};$$

$$P_{0,j} = Pr\{\tilde{L}_{n+1} = j, M_n = c | \tilde{L}_n = 0\}, \text{ for } j \geq 1;$$

$$P_{i,0} = \sum_{k=i}^{c} Pr\{\tilde{L}_{n+1} = 0, M_n = k | \tilde{L}_n = i\}, \text{ for } i \geq 1;$$

$$P_{i,j} = Pr\{\tilde{L}_{n+1} = j, M_n = c | \tilde{L}_n = i\}, \text{ for } i \geq 1, j \geq 1.$$

Then we calculate $Pr\{\tilde{L}_{n+1} = j, M_n = k | \tilde{L}_n = i\}$ as follows:

$$\begin{aligned} &Pr\{\tilde{L}_{n+1} = 0, M_n = k | \tilde{L}_n = 0\} \\ &\quad = Pr\{H_l < 1, l = 1,2, \ldots, k-1, \text{ and } H_k > 1\} \\ &\quad = \prod_{l=1}^{k-1} Pr\{H_l < 1\} \cdot Pr\{H_k > 1\} \\ &\quad = (1 - e^{-Q})^{k-1} e^{-Q}, 1 \leq k \leq c, \end{aligned}$$

where H_l $(l = 1,2, \ldots, k)$ is the exponential headway following the l-th bus arrival in the cycle; Similarly,

$$Pr\{\tilde{L}_{n+1} = j, M_n = c | \tilde{L}_n = 0\} = (1 - e^{-Q})^{c-1} \cdot \frac{e^{-Q} Q^j}{j!}, \text{ for } j \geq 1;$$

$Pr\{\tilde{L}_{n+1} = 0, M_n = k | \tilde{L}_n = i\} = (1 - e^{-Q})^{k-i} e^{-Q}$, for $1 \le i \le c-1, i \le k \le c$;

$Pr\{\tilde{L}_{n+1} = j, M_n = c | \tilde{L}_n = i\} = (1 - e^{-Q})^{c-i} \cdot \frac{e^{-Q} Q^j}{j!}$, for $1 \le i \le c-1, j \ge 1$;

$Pr\{\tilde{L}_{n+1} = j, M_n = c | \tilde{L}_n = i\} = \frac{e^{-Q} Q^{j-i+c}}{(j-i+c)!}$, for $i \ge c, j \ge i-c$;

for all other combinations of i, j, and k, $Pr\{\tilde{L}_{n+1} = j, M_n = k | \tilde{L}_n = i\} = 0$.

If we let $q_m = \frac{e^{-Q} Q^m}{m!}$, for $m = 0, 1, 2, \ldots$, then $P_{i,j}$ can be written as follows:

$P_{i,j} = q_{j-i+c}$, for $i \ge c$, and $j \ge i-c$;

$P_{i,j} = 0$, for $i \ge c$, and $j < i-c$;

$P_{i,0} = 1 - (1 - q_0)^{c+1-i}$, for $1 \le i \le c-1$;

$P_{i,j} = (1 - q_0)^{c-i} q_j$, for $1 \le i \le c-1$, and $j \ge 1$;

$P_{0,j} = P_{1,j}$, for $j \ge 0$. ■

D.2 M/D/c/SRL – Solution to the Balance Equation

From the transition probabilities given in the end of Appendix D.1, the balance equations can be written as:

$$\pi_k = \begin{cases} \pi_0(1 - (1 - q_0)^c) + \sum_{i=1}^{c-1} \pi_i (1 - (1 - q_0)^{c+1-i}) + \pi_c q_0, & \text{for } k = 0 \\ \pi_0 (1 - q_0)^{c-1} q_k + \sum_{i=1}^{c-1} \pi_i (1 - q_0)^{c-i} q_k + \sum_{i=c}^{k+c} \pi_i q_{k+c-i}, & \text{for } k = 1, 2, \ldots \end{cases}.$$

We first define:

$$\tilde{q}(z) = \sum_{k=0}^{\infty} z^k q_k = \sum_{k=0}^{\infty} z^k \frac{e^{-Q} Q^k}{k!}$$

$$= e^{Q(z-1)} \cdot \sum_{k=0}^{\infty} \frac{e^{-Qz} (Qz)^k}{k!} = e^{Q(z-1)},$$

then multiply z^k to both sides of the balance equations, and add them together. Thus we have:

$$\tilde{\pi}(z) = \pi_0 [1 + (1 - q_0)^{c-1} (\tilde{q}(z) - 1)] + \sum_{i=1}^{c-1} \pi_i \left[1 + (1 - q_0)^{c-i} (\tilde{q}(z) - 1)\right]$$

$$+ \sum_{k=0}^{\infty} z^k \sum_{i=c}^{k+c} \pi_i q_{k+c-i}.$$

For the last term of the above equation, swap the summation subscripts, so that:

$\sum_{k=0}^{\infty} z^k \sum_{i=c}^{k+c} \pi_i q_{k+c-i}$

$$= \sum_{i=c}^{\infty} \sum_{k=i-c}^{\infty} \pi_i q_{k+c-i} z^k$$

$$= \sum_{i=c}^{\infty} \sum_{j=0}^{\infty} \pi_i q_j z^{i+j-c} \qquad (\text{Let } j = k + c - i)$$

$$= z^{-c}\left(\sum_{i=c}^{\infty} \pi_i z^i\right)\left(\sum_{j=0}^{\infty} q_j z^j\right)$$

$$= z^{-c}\left(\tilde{\pi}(z) - \sum_{i=0}^{c-1} \pi_i z^i\right)\tilde{q}(z).$$

Hence,

$$\tilde{\pi}(z) = \pi_0[1 + (1 - q_0)^{c-1}(\tilde{q}(z) - 1)] + \sum_{i=1}^{c-1} \pi_i\left[1 + (1 - q_0)^{c-i}(\tilde{q}(z) - 1)\right]$$

$$+z^{-c}\left(\tilde{\pi}(z) - \sum_{i=0}^{c-1} \pi_i z^i\right)\tilde{q}(z).$$

So,

$$\tilde{\pi}(z) = \frac{\pi_0\left[1+(1-q_0)^{c-1}\left(e^{Q(z-1)}-1\right)-z^{-c}e^{Q(z-1)}\right]+\sum_{i=1}^{c-1}\pi_i\left[1+(1-q_0)^{c-i}\left(e^{Q(z-1)}-1\right)-z^{i-c}e^{Q(z-1)}\right]}{1-z^{-c}e^{Q(z-1)}}. \tag{D.2.1}$$

We need c equations to solve for c unknowns: $\pi_i (i = 0,1, \ldots, c - 1)$ in (D.2.1). By applying Rouché's theorem (Crommelin, 1932), we show that the denominator $1 - z^{-c}e^{Q(z-1)}$ has exactly c zeros in $|z| \le 1$ (see Section D.2.1 for a proof). Note that by definition, $\tilde{\pi}(z)$ converges for any z such that $|z| \le 1$, so these c zeros must be the zeros of the numerator as well. One of them is $z_0 = 1$, which is obviously a zero in the numerator; the other $c - 1$ zeros are:

$$z_k = -\frac{c}{Q} \cdot \text{LambertW}\left(-\frac{Q}{c} e^{-\frac{Q}{c}+i\frac{2k\pi}{c}}\right), \text{for } k = 1, 2, \ldots, c - 1,$$

where the function LambertW$(\cdot)$ is the inverse function of $\varphi(w) = we^w$. We solve for the $c - 1$ zeros numerically and insert them into the numerator, and obtain $c - 1$ equations:

$$\pi_0\left[1 + (1 - q_0)^{c-1}\left(e^{Q(z_k-1)} - 1\right) - {z_k}^{-c} e^{Q(z_k-1)}\right]$$

$$+\sum_{i=1}^{c-1} \pi_i\left[1 + (1 - q_0)^{c-i}(e^{Q(z_k-1)} - 1) - {z_k}^{i-c} e^{Q(z_k-1)}\right] = 0, \quad \text{for } k = 1, 2, \ldots, c - 1.$$

The last equation comes from taking the derivative of (D.2.1) with respect to z, and letting $z = 1$ (note that $\tilde{\pi}(1) = 1$):

$$\pi_0[Q(1 - q_0)^{c-1} + c - Q] + \sum_{i=1}^{c-1} \pi_i\left[Q(1 - q_0)^{c-i} + c - i - Q\right] = c - Q.$$

With these c linear equations, it is easy to solve $\pi_i (i = 0,1, \ldots, c - 1)$ using numerical tools. With the first c limiting probability values known, there are two ways to obtain the entire limiting probability distribution: (i) since $\tilde{\pi}(z)$ is determined by (D.2.1), the remaining limiting probabilities can be computed as

$$\pi_k = \frac{1}{k!}\frac{\mathrm{d}^k\tilde{\pi}(z)}{\mathrm{d}z^k}\bigg|_{z=0}, \text{ for } k = c, c+1, \ldots;$$

or (ii) we can solve for $\pi_c, \pi_{c+1}, \ldots$ in turn by iteratively applying the balance equations given at the beginning of this section. ■

D.2.1 Proof that $1 - z^{-c}e^{Q(z-1)}$ has Exactly c Zeros Inside of, or on the Unit Circle of the Complex Plane

It suffices to show that

$$z^c - e^{Q(z-1)} = 0$$

has exactly c zeros inside of, or on the unit circle.

Let $f(z) = z^c$, $g(z) = e^{Q(z-1)}$, we want to show that for any $p > 1$ and close enough to 1, $|g(z)| < |f(z)|$ for all z such that $|z| = p$.

We first show that for any $p > 1$ and close enough to 1,

$$g(p) = e^{Q(p-1)} < p^c = f(p).$$

This is because $g(1) = f(1) = 1$, and $g'(1) = Q < c = f'(1)$, where $Q < c$ is a necessary condition for a queueing system to operate in steady state.

Now let $z = p(\cos\alpha + i\sin\alpha)$,

$$|g(z)| = \left|e^{Q(p(\cos\alpha + i\sin\alpha)-1)}\right|$$

$$= \left|e^{Q(p\cos\alpha - 1)}\right| \leq e^{Q(p-1)} < p^c = |f(z)|.$$

Since $f(z)$ and $g(z)$ are both holomorphic, Rouché's theorem (Crommelin, 1932) tells us that f and $f - g$ have the same number of zeros inside the circle $|z| = p$, counting multiplicities. Given that p can be arbitrarily close to 1, we conclude that f and $f - g$ have the same number of zeros on and inside of the unit circle $|z| = 1$. Obviously $f(z)$ has exactly c zeros on and inside of the unit circle ($z = 0$ with multiplicity c), so $f(z) - g(z) = z^c - e^{Q(z-1)}$ has c zeros as well. ■

D.3 M/D/c/SRL – Average Bus Delay

When bus service time is deterministic, $\overline{W}_b = 0$, thus we only need to calculate $\overline{W}_q$.

From Appendix D.1, we have:

$Pr\{\tilde{L}_n = 0, \tilde{L}_{n+1} = 0, M_n = k\} = \pi_0(1 - e^{-Q})^{k-1}e^{-Q}$, for $k = 1,2, \dots c$;

$Pr\{\tilde{L}_n = i, \tilde{L}_{n+1} = 0, M_n = k\} = \pi_i(1 - e^{-Q})^{k-i}e^{-Q}$, for $k = i, i + 1, \dots c$, $1 \le i \le c$;

$Pr\{\tilde{L}_n = i, \tilde{L}_{n+1} = j, M_n = c\} = \pi_i P_{i,j}$, for $j \ge 1$.

By inserting the above into Equation (3.3) in Section 3.2.4, we obtain:

$$\bar{L}_q = \frac{\frac{1}{2}\pi_0 Q^2\left(1-e^{-Q}\right)^{c-1}+\frac{1}{2}\sum_{i=1}^{c-1}\pi_i Q^2\left(1-e^{-Q}\right)^{c-i}+\frac{1}{2}\sum_{i=c}^{\infty}\pi_i\left[Q^2+2Q(i-c)\right]}{Q+\pi_0\left[e^Q-Q+(Q+1-e^Q)\left(1-e^{-Q}\right)^{c-1}\right]-\sum_{i=1}^{c-1}\pi_i\left(Q+1-e^Q\right)\left(1-\left(1-e^{-Q}\right)^{c-i}\right)}.$$

Hence,

$$\overline{W} = \overline{W}_q = \frac{\bar{L}_q}{Q}$$

$$= \frac{\frac{1}{2}\pi_0 Q\left(1-e^{-Q}\right)^{c-1}+\frac{1}{2}\sum_{i=1}^{c-1}\pi_i Q\left(1-e^{-Q}\right)^{c-i}+\frac{1}{2}\sum_{i=c}^{\infty}\pi_i(Q-2c)+\sum_{i=c}^{\infty} i\pi_i}{Q+\pi_0\left[e^Q-Q+(Q+1-e^Q)\left(1-e^{-Q}\right)^{c-1}\right]-\sum_{i=1}^{c-1}\pi_i\left(Q+1-e^Q\right)\left(1-\left(1-e^{-Q}\right)^{c-i}\right)}$$

$$= \frac{\frac{1}{2}\pi_0 Q\left(1-e^{-Q}\right)^{c-1}+\frac{1}{2}\sum_{i=1}^{c-1}\pi_i Q\left(1-e^{-Q}\right)^{c-i}+\frac{1}{2}(Q-2c)\left(1-\sum_{i=0}^{c-1}\pi_i\right)+\left(\left.\frac{d\tilde{\pi}(z)}{dz}\right|_{z=1}-\sum_{i=1}^{c-1} i\pi_i\right)}{Q+\pi_0\left[e^Q-Q+(Q+1-e^Q)\left(1-e^{-Q}\right)^{c-1}\right]-\sum_{i=1}^{c-1}\pi_i\left(Q+1-e^Q\right)\left(1-\left(1-e^{-Q}\right)^{c-i}\right)}.$$

In the last equation, we eliminate π_i $(i \ge c)$ to avoid computing the numerous limiting probabilities. ■

D.4 M/G/2/SRL – Transition Probabilities

As in Section D.1, we have:

$P_{0,0} = P_{1,0} = Pr\{\tilde{L}_{n+1} = 0, M_n = 1|\tilde{L}_n = 0\} + Pr\{\tilde{L}_{n+1} = 0, M_n = 2|\tilde{L}_n = 0\}$

$= Pr\{\tilde{L}_{n+1} = 0, M_n = 1|\tilde{L}_n = 1\} + Pr\{\tilde{L}_{n+1} = 0, M_n = 2|\tilde{L}_n = 1\}$;

$P_{0,j} = P_{1,j} = Pr\{\tilde{L}_{n+1} = j, M_n = 2|\tilde{L}_n = 0\} = Pr\{\tilde{L}_{n+1} = j, M_n = 2|\tilde{L}_n = 1\}$, for $j > 0$;

$P_{i,j} = Pr\{\tilde{L}_{n+1} = j, M_n = 2|\tilde{L}_n = i\}$, for $i \ge 2$, and $j \ge i - 2$;

$P_{i,j} = 0$, otherwise.

Further, we have:

$$Pr\{\tilde{L}_{n+1} = 0, M_n = 1 | \tilde{L}_n = 1\} = Pr\{ H_1 > S_1\} = \int_{t=0}^{\infty} e^{-Qt} \mathrm{d}F_S(t),$$

where H_1 is the headway following the first (and the only) bus arrival in the cycle; and S_1 is that bus's service time.

Given that $H_1 < S_1$, there would be at least 2 arrivals in the cycle. Let τ be the time between the 2nd arrival and its departure in this cycle:

$$\tau = \max\{S_1 - H_1, S_2 | H_1 < S_1\},$$

where S_2 is the service time of the 2nd bus to arrive in the cycle.

We can derive the CDF of τ as:

$$F_\tau(t) = Pr\{ \tau < t\} = Pr\{\max\{S_1 - H_1, S_2 | H_1 < S_1\} < t\}$$

$$= \frac{Pr\{H_1 < S_1 < H_1 + t\}}{Pr\{H_1 < S_1\}} \cdot Pr\{S_2 < t\}$$

$$= \frac{\int_{h=0}^{\infty} (F_S(h+t) - F_S(h)) Q e^{-Qh} \mathrm{d}h}{Pr\{H_1 < S_1\}} \cdot F_S(t),$$

where the third equality holds because H_1, S_1, and S_2 are mutually independent.

Thus for $j \geq 0$,

$$Pr\{\tilde{L}_{n+1} = j, M_n = 2 | \tilde{L}_n = 1\} = Pr\{ H_1 < S_1\} \cdot \int_{t=0}^{\infty} \frac{e^{-Qt}(Qt)^j}{j!} \mathrm{d}F_\tau(t)$$

$$= \int_{t=0}^{\infty} \frac{e^{-Qt}(Qt)^j}{j!} \mathrm{d}\left[F_S(t) \cdot \int_{h=0}^{\infty} (F_S(h+t) - F_S(h)) Q e^{-Qh} \mathrm{d}h\right].$$

Finally, the CDF of the platoon service time of 2 buses entering the stop simultaneously would be ${F_S}^2(t)$. Then for $i \geq 2$, and $j \geq i - 2$:

$$Pr\{\tilde{L}_{n+1} = j, M_n = 2 | \tilde{L}_n = i\} = \int_{t=0}^{\infty} \frac{e^{-Qt}(Qt)^{j-i+2}}{(j-i+2)!} \mathrm{d}{F_S}^2(t).$$

In summary, we have:

$$P_{0,0} = P_{1,0} = \int_{t=0}^{\infty} e^{-Qt} \mathrm{d}\left[F_S(t) \cdot \left(1 + \int_{h=0}^{\infty} (F_S(h+t) - F_S(h)) Q e^{-Qh} \mathrm{d}h\right)\right];$$

$$P_{0,j} = P_{1,j} = \int_{t=0}^{\infty} \frac{e^{-Qt}(Qt)^j}{j!} \mathrm{d}\left[F_S(t) \cdot \int_{h=0}^{\infty} (F_S(h+t) - F_S(h)) Q e^{-Qh} \mathrm{d}h\right] \text{, for } j > 0;$$

$$P_{i,j} = \int_{t=0}^{\infty} \frac{e^{-Qt}(Qt)^{j-i+2}}{(j-i+2)!} \mathrm{d}{F_S}^2(t) \text{, for } i \geq 2, \text{ and } j \geq i - 2;$$

$P_{i,j} = 0$, otherwise. ■

D.5 M/G/2/SRL – Solution to the Balance Equation

Similarly, from the transition probabilities given at the end of Appendix D.4, we can write the balance equation in the z-domain as:

$$\tilde{\pi}(z) = (\pi_0 + \pi_1) \sum_{j=0}^{\infty} P_{0,j} z^j + \sum_{i=2}^{\infty} \pi_i \cdot \sum_{j=i-2}^{\infty} P_{i,j} z^j.$$

Let

$$G(t) = F_S(t) \cdot \int_{h=0}^{\infty} \big(F_S(h+t) - F_S(h)\big) Q e^{-Qh} \mathrm{d}h,$$

we have:

$$\sum_{j=0}^{\infty} P_{0,j} z^j = P_{0,0} + \sum_{j=1}^{\infty} P_{0,j} z^j$$

$$= Pr\{M_n = 1 | \tilde{L}_n = 1\} + Pr\{\tilde{L}_{n+1} = 0, M_n = 2 | \tilde{L}_n = 1\}$$

$$+ \sum_{j=1}^{\infty} Pr\{\tilde{L}_{n+1} = j, M_n = 2 | \tilde{L}_n = 1\} z^j$$

$$= \int_{t=0}^{\infty} e^{-Qt} \mathrm{d}F_S(t) + \sum_{j=0}^{\infty} z^j \int_{t=0}^{\infty} \frac{e^{-Qt}(Qt)^j}{j!} \mathrm{d}G(t)$$

$$= \int_{t=0}^{\infty} e^{-Qt} \mathrm{d}F_S(t) + \int_{t=0}^{\infty} e^{-Qt} \left(\sum_{j=0}^{\infty} \frac{(Qt)^j}{j!} z^j \right) \mathrm{d}G(t)$$

$$= \int_{t=0}^{\infty} e^{-Qt} \mathrm{d}F_S(t) + \int_{t=0}^{\infty} e^{Qt(z-1)} \mathrm{d}G(t).$$

And,

$$\sum_{i=2}^{\infty} \pi_i \cdot \sum_{j=i-2}^{\infty} P_{i,j} z^j$$

$$= \sum_{i=2}^{\infty} \pi_i \cdot \sum_{j=i-2}^{\infty} z^j \int_{t=0}^{\infty} \frac{e^{-Qt}(Qt)^{j-i+2}}{(j-i+2)!} \mathrm{d}F_S{}^2(t)$$

$$= \sum_{i=2}^{\infty} \pi_i \cdot \sum_{k=0}^{\infty} z^{k+i-2} \int_{t=0}^{\infty} \frac{e^{-Qt}(Qt)^k}{k!} \mathrm{d}F_S{}^2(t) \quad (\text{let } k = j - i + 2)$$

$$= z^{-2} \sum_{i=2}^{\infty} (\pi_i z^i) \cdot \int_{t=0}^{\infty} e^{-Qt} \left(\sum_{k=0}^{\infty} z^k \frac{(Qt)^k}{k!} \right) \mathrm{d}F_S{}^2(t)$$

$$= z^{-2} (\tilde{\pi}(z) - \pi_0 - \pi_1 z) \cdot \int_{t=0}^{\infty} e^{Qt(z-1)} \mathrm{d}F_S{}^2(t).$$

So,

$$\tilde{\pi}(z) = \frac{(\pi_0+\pi_1)\left(\int_{t=0}^{\infty} e^{-Qt}\mathrm{d}F_S(t)+\int_{t=0}^{\infty} e^{Qt(z-1)}\mathrm{d}G(t)\right)-z^{-2}(\pi_0+\pi_1 z)\int_{t=0}^{\infty} e^{Qt(z-1)}\mathrm{d}{F_S}^2(t)}{1-z^{-2}\int_{t=0}^{\infty} e^{Qt(z-1)}\mathrm{d}{F_S}^2(t)}. \qquad \text{(D.5.1)}$$

For any given $F_S(t)$, we can follow the steps introduced in Appendix D.2, i.e., we (i) numerically compute a root of the denominator in the region $|z| \leq 1$; (ii) insert the root into the numerator and let it be zero to obtain an equation for the unknowns π_0 and π_1; (iii) take derivative of (D.5.1) and let $z = 1$ to get a second equation; and (iv) solve these equations for π_0 and π_1 to obtain the limiting distribution $\tilde{\pi}(z)$. ■

D.6 M/G/2/SRL – Average Bus Delay

We first calculate $\overline{W}_q$.

From Appendix D.4, we know that:

$$Pr\{\tilde{L}_{n+1} = 0, M_n = 1|\tilde{L}_n = 0 \text{ or } 1\} = \int_{t=0}^{\infty} e^{-Qt}\mathrm{d}F_S(t);$$

$$Pr\{\tilde{L}_{n+1} = j \geq 0, M_n = 2|\tilde{L}_n = 0 \text{ or } 1\} = \int_{t=0}^{\infty} \frac{e^{-Qt}(Qt)^j}{j!}\mathrm{d}G(t);$$

$$Pr\{\tilde{L}_{n+1} = j \geq i-2, M_n = 2|\tilde{L}_n = i \geq 2\} = \int_{t=0}^{\infty} \frac{e^{-Qt}(Qt)^{j-i+2}}{(j-i+2)!}\mathrm{d}{F_S}^2(t).$$

So,

$$\overline{TL} = \sum_{i,j,k} Pr\{\tilde{L}_n = i, \tilde{L}_{n+1} = j, M_n = k\} \cdot TL_n$$

$$= \sum_{j=0}^{\infty} (\pi_0+\pi_1)\int_{t=0}^{\infty} \frac{e^{-Qt}(Qt)^j}{j!}\mathrm{d}G(t)\frac{j(j-1)}{2}$$

$$+\sum_{i=2}^{\infty} \pi_i \sum_{j=i-2}^{\infty}\int_{t=0}^{\infty} \frac{e^{-Qt}(Qt)^{j-i+2}}{(j-i+2)!}\mathrm{d}{F_S}^2(t)\frac{(i+j-3)(j-i+2)}{2}$$

$$= (\pi_0+\pi_1)\int_{t=0}^{\infty} e^{-Qt}\left[\sum_{j=0}^{\infty}\frac{(Qt)^j}{j!}\cdot\frac{j(j-1)}{2}\right]\mathrm{d}G(t)$$

$$+\sum_{i=2}^{\infty} \pi_i \int_{t=0}^{\infty} e^{-Qt}\left[\sum_{k=0}^{\infty}\frac{(Qt)^k}{k!}\cdot\left(\frac{k(k-1)}{2}+(i-2)k\right)\right]\mathrm{d}{F_S}^2(t) \qquad (\text{let } k = j-i+2)$$

$$= (\pi_0+\pi_1)\int_{t=0}^{\infty} e^{-Qt}\left[\sum_{j=0}^{\infty}\frac{(Qt)^j}{j!}\cdot\frac{j(j-1)}{2}\right]\mathrm{d}G(t)$$

$$+\sum_{i=2}^{\infty} \pi_i \int_{t=0}^{\infty} e^{-Qt}\left[\sum_{k=0}^{\infty}\frac{(Qt)^k}{k!}\cdot\left(\frac{k(k-1)}{2}+(i-2)k\right)\right]\mathrm{d}{F_S}^2(t)$$

$$= (\pi_0+\pi_1)\int_{t=0}^{\infty}\frac{(Qt)^2}{2}\mathrm{d}G(t)+\sum_{i=2}^{\infty}\pi_i\int_{t=0}^{\infty}\left[\frac{(Qt)^2}{2}+(i-2)Qt\right]\mathrm{d}{F_S}^2(t);$$

$\bar{A} = \sum_{i,j,k} Pr\{\tilde{L}_n = i, \tilde{L}_{n+1} = j, M_n = k\} \cdot A_n$

$= \pi_0 \int_{t=0}^{\infty} e^{-Qt} \mathrm{d}F_S(t) \cdot 1 + \pi_0 \sum_{j=0}^{\infty} \int_{t=0}^{\infty} \frac{e^{-Qt}(Qt)^j}{j!} \mathrm{d}G(t)\,(j+2) + \pi_1 \cdot$

$\sum_{j=0}^{\infty} \int_{t=0}^{\infty} \frac{e^{-Qt}(Qt)^j}{j!} \mathrm{d}G(t)\,(j+1) + \sum_{i=2}^{\infty} \pi_i \sum_{j=i-2}^{\infty} \int_{t=0}^{\infty} \frac{e^{-Qt}(Qt)^{j-i+2}}{(j-i+2)!} \mathrm{d}{F_S}^2(t)\,(j-i+2)$

$= \pi_0 \int_{t=0}^{\infty} e^{-Qt} \mathrm{d}F_S(t) + \pi_0 \int_{t=0}^{\infty} (Qt+2) \mathrm{d}G(t) +$

$\pi_1 \int_{t=0}^{\infty} (Qt+1) \mathrm{d}G(t) + \sum_{i=2}^{\infty} \pi_i \int_{t=0}^{\infty} Qt \mathrm{d}{F_S}^2(t).$

According to Equation (3.3) in Section 3.2.4,

$\bar{W}_q = \frac{\bar{L}_q}{Q} = \frac{\overline{TL}}{Q\bar{A}},$

where $\overline{TL}$ and $\bar{A}$ are given above.

We then calculate $\bar{W}_b$ in a similar way. Let TW_b be the total wait time in berths for all buses served in a cycle.

$E[TW_b | \tilde{L}_n = 0 \text{ or } 1]$

$= E[TW_b | H_1 > S_1, \tilde{L}_n = 0 \text{ or } 1] \cdot Pr\{ H_1 > S_1\}$

$+ E[TW_b | H_1 < S_1, \tilde{L}_n = 0 \text{ or } 1] \cdot Pr\{ H_1 < S_1\}$

$= E[0 | H_1 > S_1] \cdot Pr\{ H_1 > S_1\} + E[\max\{0, S_1 - H_1 - S_2\} | H_1 < S_1] \cdot Pr\{ H_1 < S_1\}$

$= E[\max\{0, S_1 - H_1 - S_2\} | H_1 < S_1] \cdot Pr\{ H_1 < S_1\}$

$= E[0 | H_1 < S_1 < H_1 + S_2] \cdot Pr\{ H_1 < S_1 < H_1 + S_2\}$

$+ E[S_1 - H_1 - S_2 | S_1 > H_1 + S_2] \cdot Pr\{ S_1 > H_1 + S_2\}$

$= \frac{\int_{h=0}^{\infty} \left[\int_{s=0}^{\infty} \left(\int_{t=0}^{\infty} t \mathrm{d}F_S(h+s+t)\right) \mathrm{d}F_S(s)\right] Qe^{-Qh} \mathrm{d}h}{Pr\{ S_1 > H_1 + S_2\}} \cdot Pr\{ S_1 > H_1 + S_2\}$

$= \int_{h=0}^{\infty} \left[\int_{s=0}^{\infty} \left(\int_{t=0}^{\infty} t \mathrm{d}F_S(h+s+t)\right) \mathrm{d}F_S(s)\right] Qe^{-Qh} \mathrm{d}h,$

where H_1, S_1, and S_2 are defined in Appendix D.4.

$E[TW_b | \tilde{L}_n \geq 2] = \frac{1}{2} E[TW_b | \tilde{L}_n \geq 2, S_1 > S_2]$ (by symmetry)

$= \frac{1}{2} E[S_1 - S_2 | S_1 > S_2]$

$$= \frac{1}{2} \frac{\int_{s=0}^{\infty}\left(\int_{t=0}^{\infty} t\mathrm{d}F_S(s+t)\right)\mathrm{d}F_S(s)}{Pr\{ S_1 > S_2\}}$$

$$= \int_{s=0}^{\infty}\left(\int_{t=0}^{\infty} t\mathrm{d}F_S(s+t)\right)\mathrm{d}F_S(s).$$

So,

$$\bar{W}_b = \frac{\overline{TW}_b}{E[M_n]} = \frac{\overline{TW}_b}{\bar{A}}$$

$$= \frac{(\pi_0+\pi_1)E[TW_b|\tilde{L}_n=0 \text{ or } 1]+(1-\pi_0-\pi_1)E[TW_b|\tilde{L}_n\geq 2]}{\bar{A}},$$

where $E[TW_b|\tilde{L}_n = 0 \text{ or } 1]$, $E[TW_b|\tilde{L}_n \geq 2]$, and $\bar{A}$ are given above. The second equality holds because the total number of buses served over a long period equals the total number of arrivals over that same period.

Therefore,

$$\bar{W} = \bar{W}_q + \bar{W}_b = \frac{\overline{TL}/Q+(\pi_0+\pi_1)E[TW_b|\tilde{L}_n=0 \text{ or } 1]+(1-\pi_0-\pi_1)E[TW_b|\tilde{L}_n\geq 2]}{\bar{A}}. \qquad \blacksquare$$