Expected transit delays in a simple system with parallel routes

Brian Fralix, Mark Holmes, Andreas Löpker

September 20, 2025

Abstract

We study a model for delays experienced by traffic along ℓ lanes when there is no opportunity to switch between lanes. A controller (who does not know the state of the system) may direct vehicles to different lanes based on their velocities.

We examine properties of the long run delay rate (as a function of the arrival rate and the distribution of velocities) for a single lane. In the multiple lane setting we compare the long run delay rates arising from different routing schemes and in particular prove that optimal partition routing (where cars are assigned to lanes based entirely on their velocities) gives smaller long run delay rate than the optimal simple random routing. Many proofs rely on coupling.

1 Introduction

Traffic congestion affects many facets of society. Commuters, emergency services, and indeed all road users can experience delays, with outcomes ranging from minor inconvenience to major adverse health impacts. Other undesirable outcomes such as noise and air pollution affect even non-road users. National and local governments design and manage traffic networks in order to alleviate traffic congestion and its consequences. This design and management can take many forms including disincentives (such as tolls), incentives for alternatives (such as improving public transport or cycling options) or changing the road network itself.

Probabilistic models for traffic flow and congestion have been around for decades see e.g. Breiman [Bri62, Bri63], Rényi [Ré64], Hawkes [Haw66, Haw68] and Zeephongsekul [Zee]. Queueing theory, a subfield of applied probability and stochastic operations research, concerns the modelling (and rigorous analysis of those models) of queuing and service systems, that typically also experience congestion. There have been several more recent works in the literature utilizing queuing theory in the context of traffic flow modelling or analysis, including Heidemann [Hei94] and Jain and Smith [JS97]. We refer readers to the survey by Van Woensel and Vandaele [VWV07] for further references prior to 2007. Most relevant to our work is the work of Chao et al. [CHR15]. Therein the authors introduce and study the 'tollbooth tandem queue', which is a type of infinite-server queue where the servers are arranged in series (instead of in parallel). When a

customer/job arrives at the system, its service commences immediately, but that job cannot leave until all prior arrivals have left the system. The authors of [CHR15] point out that the tollbooth tandem queue can be interpreted as a single-lane traffic model, but the main objective of [CHR15] is to derive various distributions and performance measures associated with this queueing system.

We seek to understand how arriving cars should be routed to lanes in this traffic model when there are multiple lanes and passing within a lane is not allowed. More precisely, we are interested in reducing or minimising the *delay per unit time* (which we will define precisely below) occurring in the system. Under the routing schemes we consider, the traffic flows in different lanes will be independent of each other. Therefore a substantial part of our effort will be directed to investigating properties of the delay per unit time in a single lane.

The model studied here is very much a 'toy model', but we hope that our treatment will encourage both: (i) further study of this and similar toy models for traffic congestion; and (ii) extensions to more realistic models (and perhaps even practical recommendations).

2 Main Results

Suppose there are $\ell \in \mathbb{N}$ lanes on a stretch of highway, labelled $1, \ldots, \ell$, all of unit length, enabling travel from a common source/routing point to a common destination. Cars arriving at the source are immediately assigned to a lane, and cannot change lanes at any point along the highway. They can, therefore, be delayed by slower cars travelling in front of them. This raises a very general question: how should arriving cars be assigned to lanes in order to minimise delays? We will address this question for some specific classes of routing schemes.

Cars arrive at the source in accordance to a Poisson process $\{N(t); t \geq 0\}$ having rate λ, with $N(t)$ denoting the number of cars who enter the highway within the time interval $[0, t]$, $t > 0$. The i-th car arrives at time T_i, travelling at speed V_i: the *workload* associated to this car is $W_i := 1/V_i \sim F$, where W_i can be interpreted as the amount of time it takes the i-th car to move from the routing point to the endpoint of the highway, when it is unimpeded by any other cars. We assume throughout that $\{W_i\}_{i\geq 0}$ is an i.i.d. sequence, independent of the arrival process, with cdf F satisfying $F(0-) := \lim_{x\uparrow 0} F(x) = 0$, as well as $F(\infty) := \mathbb{P}(W < \infty) = 1$. Let $\mathrm{supp}(F)$ denote the support of F, where we recall $t \in \mathrm{supp}(F)$ if, for each $\epsilon > 0$, $F(t+\epsilon) - F(t-\epsilon) > 0$.

In a real system an arriving car may choose (or be routed to) a lane depending on what is in front of it at the time of arrival (e.g. based on the speed of or distance to the closest car in each lane). However we will restrict our attention to routing schemes in which each arriving car is routed to lane $j \in [\ell] := \{1, 2, \ldots, \ell\}$ with probability q_j (independent of all other cars), so that $\sum_{j\in[\ell]} q_j = 1$. The lane entered could be chosen completely at random, or could depend on the velocity of the arriving car. This will be specified more precisely later.

If the n-th car is unimpeded by any slower cars encountered on the road, it arrives at the endpoint at time $D_n := T_n + W_n$. However, passing is not possible (as cars cannot switch lanes), so if the n-th arrival catches up to another car in front of it, it is impeded by that slower car. It therefore adjusts its velocity to match that of the impeding car, and they reach the destination at the same time. For each integer $n \geq 1$, let S_n denote the sojourn time of the car on the highway, and let $\Delta_n := S_n - W_n$ denote the total delay experienced by the n-th car from driving on the entire road segment. Thus, $T_n + W_n$ is

the time at which the n-th car expects to arrive to the endpoint, assuming ideal traffic conditions, and $T_n + S_n$ is the actual time this arrival arrives at the endpoint.

Next, define, for each $t \geq 0$,

$$\Delta_n(t) := ((T_n + S_n) \wedge t) - ((T_n + W_n) \wedge t) \tag{1}$$

as the amount of time the nth arrival is overdue at time t. Clearly, $\Delta_n(t) \to \Delta_n$ as $t \to \infty$. Finally, for each $t \geq 0$ we define $L_t(\lambda, F) = 0$ when $\lambda = 0$ and otherwise

$$L_t(\lambda, F) := \sum_{n=1}^{\infty} \Delta_n(t) \tag{2}$$

which represents the cumulative delay at time t of all cars arriving to the highway in $[0, t]$ (clearly $\Delta_n(t) = 0$ when $t < T_n$). We study the long-run delay rate $\lim_{t\to\infty} L_t(\lambda, F)/t$. When $\ell = 1$ we express the long-run delay rate as $\mathcal{L}(\lambda, F)$. Existence and finiteness of the limit are non-trivial, and are the subject of Theorem 1 below.

Let, for each $t \geq 0$, X_t denote the number of customers traversing the highway at time t. Observe (assuming $X_0 = 0$) that for each $t \geq 0$,

$$X_t = \sum_{n=1}^{\infty} \mathbf{1}_{\{T_n \leq t, T_n + S_n > t\}}.$$

Many properties of $\{X_t; t \geq 0\}$ play an important role in establishing our main results, e.g. when $\ell = 1$, $\{X_t; t \geq 0\}$ can be interpreted as the queue-length process of the tollbooth tandem queue of [CHR15], and this interpretation will play a role in some of our arguments.

We seek a basic understanding of how certain routing schemes compare in terms of the long run delay rate. To this end we will establish various properties (not found in [CHR15]) of $\mathcal{L}(\lambda, F)$ as a function of both λ and F, and then use these (and other methods) to demonstrate that some routing strategies are superior to others. Along the way we will also present some examples that illustrate interesting features of $\mathcal{L}(\lambda, F)$.

In order to establish part of our main limit theorem, it helps to introduce the infinite-server queue process $\{Y_t; t \geq 0\}$ associated with $\{X_t; t \geq 0\}$, where for each $t \geq 0$,

$$Y_t := \sum_{n=1}^{\infty} \mathbf{1}_{\{T_n \leq t, T_n + W_n > t\}}.$$

Note that $X_t \geq Y_t$ for each $t \geq 0$, simply because $S_n \geq W_n$ for each integer $n \geq 1$. Moreover, standard regenerative arguments show that both $\{X_t; t \geq 0\}$ and $\{Y_t; t \geq 0\}$ have a limiting distribution when $\mathbb{E}[W] < \infty$, and it can also be shown (see [CHR15]) that if X_∞ and Y_∞ denote random variables whose laws, respectively, are those of the limiting distribution of $\{X_t; t \geq 0\}$ and the limiting distribution of $\{Y_t; t \geq 0\}$, then

$$\mathbb{E}[X_\infty] = \lim_{t\to\infty} \mathbb{E}[X_t], \qquad \mathbb{E}[Y_\infty] = \lim_{t\to\infty} \mathbb{E}[Y_t].$$

Our first main result establishes the existence of the a.s. limit $\mathcal{L}(\lambda, F)$ and gives a concise formula for it.

Theorem 1. *For a single lane with arrival rate λ and workload distribution F, the almost-sure limit $\mathcal{L}(\lambda, F) \in [0, \infty]$ exists, is non-random and satisfies the following.*

(i) *$\mathcal{L}(\lambda, F) < \infty$ if and only if $\mathbb{E}[W^2] < \infty$.*

(ii) *If* $\mathbb{E}[W^2] < \infty$ *then*

$$\mathcal{L}(\lambda, F) = \lambda \int_0^\infty \left(1 - e^{-\lambda H(s)}\right) F(s)\, ds, \tag{3}$$

where $H(y) := \int_y^\infty \overline{F}(s) ds$.

Proof. We provide a simple proof of Theorem 1 for the special case where $\mathbb{E}[W] < \infty$, as it is analogous to arguments typically used in queueing theory to establish Little's law; see e.g. Whitt [Whi91, Whi92] and in particular both Serfozo [Ser94] and Chapter 5 of Serfozo [Ser99]. Proving this result for the case where $\mathbb{E}[W] = \infty$ appears to be significantly more difficult, and the reader can find the argument in Section 3.

The key to establishing this result when $\mathbb{E}[W] < \infty$ is to realise that, for each $t \geq 0$,

$$L_t(\lambda, F) = \sum_{n=1}^{\infty} [((T_n + S_n) \wedge t) - ((T_n + W_n) \wedge t)] = \int_0^t (X_s - Y_s) ds. \tag{4}$$

Note also that, as observed in [CHR15], the busy period distribution associated with $\{X_t\}_{t\geq 0}$ is equal to the busy period distribution associated with $\{Y_t\}_{t\geq 0}$, and the latter is clearly finite with probability one if and only if $\mathbb{E}[W] < \infty$. From here, it follows from standard regenerative process theory (see e.g. Theorem 3.1 on page 178 of Asmussen [Asm03]) that

$$\frac{L_t(\lambda, F)}{t} \to \mathbb{E}[X_\infty - Y_\infty]$$

almost surely as $t \to \infty$, so that $\mathcal{L}(\lambda, F)$ exists and is equal to $\mathbb{E}[X_\infty - Y_\infty]$ when $\mathbb{E}[W] < \infty$. Finally, Chao et al. show at the bottom of page 946 of [CHR15] that

$$\mathbb{E}[X_\infty - Y_\infty] = \lambda \int_0^\infty \left(1 - e^{-\lambda H(s)}\right) F(s)\, ds$$

which proves statement (ii), and completes the proof of Theorem 1 when $\mathbb{E}[W] < \infty$. It is also easy to show that when $\mathbb{E}[W] < \infty$, $\mathcal{L}(\lambda, F) < \infty$ if and only if $\mathbb{E}[W^2] < \infty$. ■

A similar expression to (3) holds for the expected value of $L_t(\lambda, F)$.

Proposition 2. *For a single lane with arrival rate* λ *and workload distribution* F*,*

$$\mathbb{E}[L_t(\lambda, F)] = \lambda \int_0^t \int_0^s \left[1 - e^{-\lambda \int_u^s \overline{F}(v) dv}\right] F(u) du ds. \tag{5}$$

Moreover,

$$\lim_{t\to\infty} \frac{\mathbb{E}[L_t(\lambda, F)]}{t} = \mathcal{L}(\lambda, F). \tag{6}$$

Formula (6) follows trivially from (5), even when $\mathbb{E}[W] = \infty$, which is interesting in light of what we found whilst proving Theorem 1.

Proof. This result follows from Formula (4). Indeed, on page 946 of [CHR15], the authors show that for each $s \geq 0$,

$$\mathbb{E}[X_s - Y_s] = \lambda \int_0^s F(u) \left[1 - e^{-\lambda \int_u^s \overline{F}(v) dv}\right] du \tag{7}$$

and after taking the expected value of both sides of (4), then applying (7), we get

$$\mathbb{E}[L_t(\lambda, F)] = \lambda \int_0^t \int_0^s F(u) \left[1 - e^{-\lambda \int_u^s \overline{F}(v)dv}\right] duds$$

proving (5). As noted above, (6) is a simple consequence of (5). ■

Example 1. Suppose that W takes the values a and $b > a$ with probabilities $1-p$ and p respectively. Then one can easily evaluate (3) to get

$$\mathcal{L}(\lambda, F) = \frac{1-p}{p}\Big[p\lambda(b-a) - 1 + e^{-\lambda p(b-a)}\Big]. \tag{8}$$

A calculus exercise shows that the value of p which maximizes $\mathcal{L}$ is always in $(0, 1/2)$. Thus we are worse off in this example having slightly more delayees than delayers.

One can easily evaluate (5) for this example too. Indeed, with $\delta = b - a$,

$$\mathbb{E}[L_t(\lambda, F)] = \begin{cases} 0, & \text{if } t \le a \\ \frac{1-p}{p}\left[\lambda p\frac{(t-a)^2}{2} - (t-a) + \frac{1-e^{-\lambda p(t-a)}}{\lambda p}\right], & \text{if } t \in (a, b], \\ \frac{1-p}{p}\left[\lambda p\frac{\delta^2}{2} - \delta + \frac{1-e^{-\lambda p\delta}}{\lambda p} + (t-b)\big[\lambda p\delta - 1 + e^{-\lambda p\delta}\big]\right], & \text{if } t > b. \end{cases}$$

Note that the growth is linear after time b (the maximum of the *bounded* support).

In view of Theorem 1(i) we make the following assumption.

Assumption A. We henceforth assume (unless otherwise stated) that $\mathbb{E}[W^2] < \infty$.

Our next result establishes many properties of $\mathcal{L}(\lambda, F)$ that will be useful when we study routing strategies for cars travelling on a multi-lane highway. The proof is in Section 3.

Theorem 3. *Suppose that* $\mathrm{supp}(F)$ *contains at least two elements.*

(a) $\mathcal{L}(\lambda, F) > 0$ *when* $\lambda \in (0, \infty)$.

(b) $\mathcal{L}(\lambda, F)$ *is strictly increasing in* λ *on* $[0, \infty)$ *and differentiable in* λ *on* $(0, \infty)$.

(c) For $\alpha, \beta \ge 0$, *if* F' *is the cdf of* $W' = \beta W + \alpha$ *then*

$$\mathcal{L}(\lambda, F') = \mathcal{L}(\beta\lambda, F). \tag{9}$$

(d) $\mathcal{L}(\lambda, F)$ *is superadditive in* λ *on* $[0, \infty)$, *i.e. for* $\lambda_1, \lambda_2 \ge 0$

$$\mathcal{L}(\lambda_1 + \lambda_2, F) \ge \mathcal{L}(\lambda_1, F) + \mathcal{L}(\lambda_2, F). \tag{10}$$

(e) If $W \sim F$ *and* $W' \sim F'$ *then*

$$\big|\mathcal{L}(\lambda, F) - \mathcal{L}(\lambda, F')\big| \le \lambda\big|\mathbb{E}[W'] - \mathbb{E}[W]\big| + \lambda^2 \int_0^\infty s\big|\overline{F'}(s) - \overline{F}(s)\big|\, ds.$$

In particular, if F *has finite variance and* $\{F_n\}$ *is a sequence that satisfies* $\int_0^\infty s|\overline{F}_n(s) - \overline{F}(s)|\, ds \to 0$ *as* $n \to \infty$, *then* $\mathcal{L}(\lambda, F_n) \to \mathcal{L}(\lambda, F)$ *as* $n \to \infty$.

(f) $\mathcal{L}(\lambda, F)$ *is convex in* λ *on* $[0, \infty)$.

Remark 2. Let $b-a=1$ in Example 1. Let G_p denote the workload cdf associated to this example (it depends on p). For small λ we have that $\mathcal{L}(\lambda, G_p) \approx \frac{\lambda^2}{2}p(1-p)$. For large λ we have that $\mathcal{L}(\lambda, G_p) \approx (1-p)\lambda$. It follows that

$$\begin{aligned} &\mathcal{L}(\lambda, G_{1/2}) > \mathcal{L}(\lambda, G_{1/3}), && \text{for } \lambda \text{ sufficiently small,} \\ &\mathcal{L}(\lambda, G_{1/2}) < \mathcal{L}(\lambda, G_{1/3}), && \text{for } \lambda \text{ sufficiently large.} \end{aligned}$$

This shows that (in general) whether one workload distribution gives larger long-run delay rate than another depends on λ.

Further results concerning a single lane appear in Section 3. In particular we will present results concerning asymptotics for small and large λ, as well as conditions under which one workload distribution gives larger long-run delay rate than another *for all* λ.

We now turn our attention to multiple lanes. Let $\mathcal{F} = \sigma(\{W_n\}_{n\in\mathbb{N}}, \{T_n\}_{n\in\mathbb{N}}, \{U_n\}_{n\in\mathbb{N}})$, where $\{U_n\}_{n\in\mathbb{N}}$ are i.i.d. standard uniform random variables that are independent of the arrival process. For fixed $\ell \geq 2$, let $\{R_n\}_{n\in\mathbb{N}}$ be any $\mathcal{F}$-measurable assignment of arrivals to lanes, where R_n denotes the lane traversed by the nth arrival. It is a trivial exercise to show that the delay experienced by any individual under lane assignments $\{R_n\}_{n\in\mathbb{N}}$ is not more than that when all arrivals are sent to lane 1.

It is interesting to study how vehicle delay accumulates over time, whenever the vehicles choose lanes according to some fixed policy. Examples of such policies include the following:

- **Random Routing**: When cars select lanes under this policy, an arriving car chooses lane j with probability q_j, independent of its speed/workload, or any other information. Standard thinning properties of Poisson processes imply that the long-run delay rate is given by

$$\mathcal{L}(\lambda, F \,|\, \boldsymbol{q}) := \lim_{t\to\infty} \frac{L_t(\lambda, F)}{t} = \sum_{j=1}^{\ell} \mathcal{L}(\lambda q_j, F).$$

  We refer to this routing scheme as the *random routing scheme*.

- **Partition Routing**: Another interesting routing scheme is the *partition routing scheme*: given a partition $\mathbf{I} := \{I_j\}_{j\in[\ell]}$ of $[0,\infty)$ consisting of Borel sets, an arriving vehicle chooses lane j if its workload is an element of I_j. Hence, an arriving vehicle chooses lane j with probability $q_j := \mathbb{P}(W \in I_j)$. Because the workload of an arriving vehicle is independent of the workloads of all other vehicles, thinning arguments can again be used to show that the long-run delay rate is

$$\mathcal{L}(\lambda, F \,|\, \mathbf{I}) := \lim_{t\to\infty} \frac{L_t(\lambda, F)}{t} = \sum_{j=1}^{\ell} \mathcal{L}(\lambda q_j, F_j)$$

  where the cdf $F_j : \mathbb{R} \to [0,1]$ is defined as $F_j(t) := \mathbb{P}(W \leq t \mid W \in I_j)$.

- **Cyclic Routing**: A third natural routing scheme that would be of interest to study for comparison purposes is the case where $R_n - 1 = n \mod \ell$. We call this *cyclic routing* and note that in this case the arrival process in each lane is not Poisson.

In this paper we primarily investigate properties of random and partition routing schemes. Although our main results in this setting are not surprising, some other intuitively reasonable statements turn out to be false in general.

The following result shows that the uniform random routing policy is optimal among the set of all random routing policies.

Proposition 4. *The optimal random routing policy* $\mathbf{q}^* := [q_i^*]_{i\in[\ell]}$ *among the set of all random routing policies is* $q_i^* = 1/\ell$, $i \in [\ell]$.

Proof. Recall from Theorem 3(f) that $\mathcal{L}(\lambda, F)$ is convex in λ. Therefore, for any random routing policy $\mathbf{q} := [q_i]_{i\in[\ell]}$ satisfying $q_i \neq 1/\ell$ for some $i \in [\ell]$, we have

$$\mathcal{L}(\lambda/\ell, F) = \mathcal{L}\left(\sum_{i=1}^{\ell} \frac{1}{\ell}\lambda q_i, F\right) \leq \sum_{i=1}^{\ell} \frac{1}{\ell}\mathcal{L}(\lambda q_i, F)$$

from which we get $\sum_{i=1}^{\ell} \mathcal{L}(\lambda/\ell, F) = \ell\mathcal{L}(\lambda/\ell, F) \leq \sum_{i=1}^{\ell} \mathcal{L}(\lambda q_i, F)$, as required. ■

We expect that cyclic routing is superior to random routing. The next result establishes the superiority of partition routing as compared to random routing, when F is continuous.

Theorem 5. *If* F *is continuous then, for any* $\lambda > 0$, *optimal partition routing has a smaller delay rate than optimal random routing.*

In general it seems to be a very difficult task to find optimal partition routing schemes for arbitrary F. A reasonable first guess is that optimal partitions put cars together with other cars of similar speed, in such a way that each lane is assigned a single interval of speeds/workloads. This turns out to not be the case in general, with 3 distinct velocities (see Example 3) already providing counterexamples.

Example 3. Consider the case $\ell = 2$, with three workloads $w_i = i$, $i = 1, 2, 3$ with probabilities $(1, 20, 1)/22$, and $\lambda = 10$. Of the three non-trivial partition rules (corresponding to which workload gets a lane of its own), the one that puts the cars with workload 2 in their own lane (hence the highest and lowest workloads in the other lane) is optimal. Moreover this partition gives lower long-run delay rate than any semi-randomized partition that puts workloads 1 and 3 in different lanes and splits the other workloads randomly between the two lanes (probability q' of being assigned to lane 1 - see Figure 1).

This is not an artifact of W being discrete, as we observe the same phenomenon with an approximating continuous W (recall Theorem 3(e)).

Following the previous example, the next theorem should now be considered non-obvious.

Theorem 6. *Suppose* $F \sim U(a, b)$. *Then the optimal partition routing scheme assigns to each lane, a single interval of length* $(b - a)/\ell$.

Open Problem 4. Give non-trivial sufficient conditions on F that guarantee that the optimal partition rule assigns to each lane a single interval.

The rest of the paper is organised as follows. In Section 3 we prove various results about the long-run delay rate for a single lane, including Theorems 1, 2, and 3. In Section 4 we prove Theorems 5 and 6.

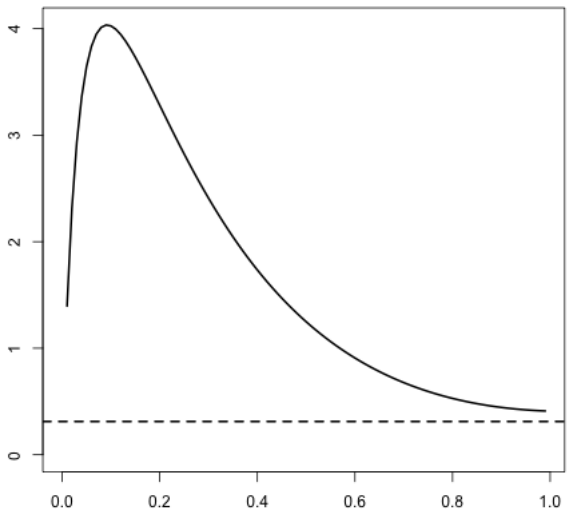

Figure 1: A plot of $\mathcal{L}$ as a function of q' in the setting of Example 3, where there are 3 distinct workloads (1,2,3) and two lanes. Workloads 1 are put in lane 1, and workloads 3 are put in lane 2. Workloads equal to 2 are assigned at random to either lane (lane 1 with probability q'). One obtains a smaller $\mathcal{L}$ (dotted line) by putting workloads equal to 2 in lane 2 and all others in lane 1.

3 Single lane proofs

In this section we finish the proof of Theorem 1, and we also prove Theorem 3. We also include some additional results (and proofs) concerning properties of $\mathcal{L}(\lambda, F)$. Throughout this section $\ell = 1$.

We start by showing that the delay rate is infinite when W has infinite expectation, thus completing the proof of Theorem 3.

Lemma 7. *If $\mathbb{E}[W] = \infty$ then $L_t(\lambda, F)/t \to \infty$ almost surely as $t \to \infty$.*

Proof. The proof is by coupling. Roughly speaking we will compare the delay up to time t with a sum of i.i.d. random variables. The proof is somewhat "technical", so before presenting the details, let us indicate the main ideas of the proof.

Consider the *second* arrival. This arrival has workload $W_2 < \infty$, so on the event that $W_2 \leq c$, the expected delay that it will experience is infinite due to the fact that the expected workload of the first arrival is infinite. The delay that it will experience up to time t is of course finite, but that delay will have large expected value if t is large (the second arrival should arrive well before time t), and $W_2 \leq c$ (which has positive probability). More generally, *even* arrivals who: (1) arrive well before time t; and (2) have workload that is not large; and (3) have the previous arrival arrive a short time before them, will be expected to be delayed a lot by that previous arrival. Roughly speaking, the proof below will combine these observations in a rigorous way, and will show that by counting only delays to even arrivals ignoring all delayers but the previous

odd arrival (call this a *paired delay*) we get a lower bound on

$$\liminf_{t\to\infty} \frac{L_t(\lambda, F)}{t},$$

that is as large as we like by choosing "well before", "not large" and "short time" appropriately in (1),(2),(3). The proof ultimately utilises the law of large numbers, so we will need to show that the paired delays above dominate an i.i.d. collection of random variables.

We now commence the proof proper. First, observe that since $\lim_{x\to\infty} F(x) = 1$ by assumption, there exists $c > 0$ such that $\mathbb{P}(W < c) > 0$. We also define, for each integer $i \geq 1$,

$$M_i := \sup_{1\leq \ell \leq i} (T_i + W_i), \qquad M_{2i}^{'} := \max(T_{2i-1} + W_{2i-1}, T_{2i} + W_{2i}).$$

In particular M_{2i} is the departure time of the $2i$-th arrival, while M'_{2i} would be the departure time of the $2i$-th arrival if this individual could only be delayed by the previous arrival. Clearly $M'_{2i} \leq M_{2i}$.

Next, define, for each $t \geq 0$, $N_e(t)$ as the number of evenly-indexed arrival times less than or equal to t, i.e.

$$N_e(t) := \sum_{k=1}^{\infty} \mathbf{1}_{\{T_{2k}\leq t\}}.$$

Given a fixed positive integer $s_0 > c + 1$, observe that for each $t > s_0$,

$$\begin{aligned} L_t(\lambda, F) &= \sum_{i=1}^{N(t)} \left[M_i \wedge t - (T_i + W_i) \wedge t\right] \\ &\geq \sum_{i=1}^{N_e(t-s_0)} \mathbf{1}_{\{W_{2i}\leq c, T_{2i}-T_{2i-1}\leq 1, M_{2i}>T_{2i}+W_{2i}\}} \left[M_{2i} \wedge t - (T_{2i} + W_{2i}) \wedge t\right]. \end{aligned} \tag{11}$$

The summand in the first line above is the delay experienced by the ith arrival up to time t. We get a lower bound on the sum by only counting even arrivals, and a smaller lower bound by only counting those even arrivals $2i$ that arrived before $t - s_0$ (so "well before" t if s_0 is not small) whose workload is not large ($\leq c$), and whose predecessor, the $2i - 1$st arrival, arrived at most 1 time unit (a "short time") before them. We also need only count those even arrivals who are actually delayed ($M_{2i} > T_{2i} + W_{2i}$). This results in the lower bound (11), which is equal to

$$\sum_{i=1}^{N_e(t-s_0)} \mathbf{1}_{\{W_{2i}\leq c, T_{2i}-T_{2i-1}\leq 1, M_{2i}>T_{2i}+W_{2i}\}} \left[M_{2i} \wedge t - (T_{2i} + W_{2i})\right]. \tag{12}$$

To see this note that on the set $\{T_{2i} \leq t - s_0, W_{2i} \leq c, T_{2i} - T_{2i-1} \leq 1, M_{2i} > T_{2i} + W_{2i}\}$,

$$T_{2i} + W_{2i} \leq t - s_0 + c = t - (s_0 - c) \leq t.$$

Moreover, since $M_{2i}^{'} \leq M_{2i}$ we have that

$$(12) \geq \sum_{i=1}^{N_e(t-s_0)} \mathbf{1}_{\{W_{2i}\leq c, T_{2i}-T_{2i-1}\leq 1, M'_{2i}>T_{2i}+W_{2i}\}} \left[M_{2i}^{'} \wedge t - (T_{2i} + W_{2i})\right].$$

Next we define, for each integer $i \geq 1$,

$$\hat{M}_{2i} := t - s_0 + W_{2i-1} - (T_{2i} - T_{2i-1}).$$

and when $T_{2i} \leq t - s_0$ and $M'_{2i} > T_{2i} + W_{2i}$, we have

$$\hat{M}_{2i} \geq T_{2i} + W_{2i-1} - (T_{2i} - T_{2i-1}) = T_{2i-1} + W_{2i-1} = M'_{2i}.$$

Our next goal is to show that

$$\begin{aligned}
&\mathbf{1}_{\{W_{2i} \leq c, T_{2i} - T_{2i-1} \leq 1, M'_{2i} > T_{2i} + W_{2i}\}} \left[M'_{2i} \wedge t - (T_{2i} + W_{2i}) \right] \\
&\geq \mathbf{1}_{\{W_{2i} \leq c, T_{2i} - T_{2i-1} \leq 1, M'_{2i} > T_{2i} + W_{2i}\}} \left[\hat{M}_{2i} \wedge t - (t - s_0 + W_{2i}) \right].
\end{aligned}$$

Observe that, when $T_{2i} \leq t - s_0, W_{2i} \leq c, T_{2i} - T_{2i-1} \leq 1, M'_{2i} > T_{2i} + W_{2i}$, we have

$$t - s_0 + W_{2i} \leq t - s_0 + c = t - (s_0 - c) < t$$

so it suffices to prove the claim for the following three cases: (a) $t \in (t - s_0 + W_{2i}, M'_{2i}]$, (b) $t \in (M'_{2i}, \hat{M}_{2i})$, and (c) $\hat{M}_{2i} \leq t$.

Assuming first that $\hat{M}_{2i} \leq t$ (i.e. case (c)), it follows that $M'_{2i} \leq t$, and

$$\begin{aligned}
\hat{M}_{2i} \wedge t - (t - s_0 + W_{2i}) &= \hat{M}_{2i} - (t - s_0 + W_{2i}) \\
&= [t - s_0 + W_{2i-1} - (T_{2i} - T_{2i-1})] - (t - s_0 + W_{2i}) \\
&= W_{2i-1} + T_{2i-1} - (W_{2i} + T_{2i}) \\
&= M'_{2i} - (T_{2i} + W_{2i}) \\
&= M'_{2i} \wedge t - (T_{2i} + W_{2i})
\end{aligned}$$

which proves the inequality for this case.

Next, suppose that $t - s_0 + W_{2i} < t \leq M'_{2i}$ (i.e. case (a)). Then $M'_{2i} \wedge t = t = \hat{M}_{2i} \wedge t$, and moreover,

$$M'_{2i} \wedge t - (T_{2i} + W_{2i}) = \hat{M}_{2i} \wedge t - (T_{2i} + W_{2i}) \geq \hat{M}_{2i} \wedge t - (t - s_0 + W_{2i})$$

which proves the inequality for this case.

Finally, suppose that $t \in (M'_{2i}, \hat{M}_{2i})$ (i.e. case (b)). In this case,

$$\begin{aligned}
\hat{M}_{2i} \wedge t - (t - s_0 + W_{2i}) &\leq \hat{M}_{2i} - (t - s_0 + W_{2i}) \\
&= W_{2i-1} + T_{2i-1} - (W_{2i} + T_{2i}) \\
&= M'_{2i} - (W_{2i} + T_{2i}) \\
&= M'_{2i} \wedge t - (W_{2i} + T_{2i})
\end{aligned}$$

and the claimed inequality holds again.

We have shown that the summand in (11) is at least

$$\mathbb{1}_{\{W_{2i} \leq c, T_{2i} - T_{2i-1} \leq 1, W_{2i-1} > W_{2i} + (T_{2i} - T_{2i-1})\}} \left[(t \wedge \hat{M}_{2i}) - (t - s_0 + W_{2i}) \right], \tag{13}$$

where we note that the last event in the indicator is the same as the event $M'_{2i} > T_{2i} + W_{2i}$. Finally, since $W_{2i} \leq c$ and $T_{2i} - T_{2i-1} \leq 1$ when the indicator is non-zero, it follows that (13) is at least

$$\mathbb{1}_{\{W_{2i} \leq c, T_{2i} - T_{2i-1} \leq 1\}} \mathbb{1}_{\{W_{2i-1} > c+1\}} \left[(t \wedge [t - s_0 + W_{2i-1} - 1]) - (t - s_0 + c) \right].$$

This is equal to

$$\mathbb{1}_{\{W_{2i}\le c, T_{2i}-T_{2i-1}\le 1\}}\mathbb{1}_{\{W_{2i-1}>c+1\}}\cdot\begin{cases} s_0-c & , \text{ if } W_{2i-1}-1\ge s_0\\ W_{2i-1}-1-c & , \text{ if } W_{2i-1}-1<s_0.\end{cases}$$
$$=\mathbb{1}_{\{W_{2i}\le c, T_{2i}-T_{2i-1}\le 1\}}\Big[\big((W_{2i-1}-(1+c))\vee 0\big)\wedge(s_0-c)\Big].$$

We have shown that

$$L_t(\lambda,F)\ge\sum_{i=1}^{N_e(t-s_0)}\mathbf{1}_{\{W_{2i}\le c, T_{2i}-T_{2i-1}\le 1\}}\Big[\big((W_{2i-1}-1-c\big)\vee 0)\wedge(s_0-c)\Big]. \tag{14}$$

The classical strong law of large numbers shows that almost surely as $n\to\infty$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{W_{2i}\le c, T_{2i}-T_{2i-1}\le 1\}}\Big[\big((W_{2i-1}-1-c\big)\vee 0)\wedge(s_0-c)\Big]\to\mathfrak{L}(s_0),$$

(since the summands are i.i.d. random variables) where

$$\mathfrak{L}(s_0)=\mathbb{E}\left[\mathbf{1}_{\{W_2\le c, T_2-T_1\le 1\}}\right]\mathbb{E}\left[\big((W_1-1-c\big)\vee 0)\wedge(s_0-c)\right].$$

Since $N_e(t-s_0)/t\to\lambda/2$ almost surely as $t\to\infty$ it follows from (14) that

$$\liminf_{t\to\infty}\frac{L_t(\lambda,F)}{t}\ge\frac{\lambda}{2}\mathfrak{L}(s_0)$$

and since $\mathbb{E}[W]=\infty$, $\mathfrak{L}(s_0)\to\infty$ as $s_0\to\infty$ which completes the proof. ■

The following proposition is elementary, but it is useful because it provides us with another interpretation of $\mathcal{L}(\lambda,F)$ when $\ell=1$.

Proposition 8. *Suppose that $\ell=1$. Then for each integer $n\ge 1$, the sojourn time random variable S_n satisfies*

$$S_n=\max_{1\le j\le n}(W_j-(T_n-T_j)).$$

Moreover, $S_n\xrightarrow{d}S_\infty$ as $n\to\infty$, where

$$S_\infty\stackrel{d}{=}\sup_{j\ge 0}(W_j-T_j)$$

and as a consequence,

$$\mathcal{L}(\lambda,F)=\lambda\mathbb{E}\left[\sup_{j\ge 0}((W_j-W_0)-T_j)\right]. \tag{15}$$

Proof. First, observe that when $\ell=1$, the departure time of the nth car must be greater than or equal to the departure time of cars $1,2,\dots,n-1$. More particularly, it is clear from the dynamics of the model that, for each integer $n\ge 1$,

$$D_n=\max_{1\le k\le n}(T_k+W_k)$$

and, therefore,

$$S_n = D_n - T_n = \max_{1\le k\le n}(W_k + T_k) - T_n = \max_{1\le k\le n}(W_k - (T_n - T_k))$$

proving the first half of the claim. The remaining claim follows from an argument that is essentially the same as that used to establish the limiting distribution of the Lindley recursion. Indeed, letting W_0 be a generic random variable having CDF F, independent of both $\{W_k\}_{k\ge1}$ and the arrival processes, we have that for each integer $n \ge 1$,

$$S_n = \max_{1\le j\le n}(W_j - (T_n - T_j)) \stackrel{d}{=} \max_{0\le j\le n-1}(W_j - T_j) =: \tilde{S}_n.$$

Clearly the sequence of random variables $\tilde{S}_n$ is nondecreasing in n, and therefore it converges almost-surely as $n \to \infty$ to

$$\tilde{S}_\infty := \sup_{j\ge0}(W_j - T_j)$$

proving $S_n \stackrel{d}{\to} S_\infty$ as $n \to \infty$. Once this has been established, Equality (15) follows from recalling $\mathcal{L}(\lambda, F) = \mathbb{E}[X_\infty - Y_\infty]$, then applying Little's law. ■

We first obtained the formula (3) in the following way, which will be useful for later results.

Lemma 9. *For $K > 0$ let $\Delta^{(K)}$ denote the total delay experienced by a marked car arriving at exactly time K. Then*

$$\lim_{K\to\infty}\mathbb{E}[\Delta^{(K)}] = \int_0^\infty \big(1 - e^{-\lambda H(s)}\big)F(s)\,ds.$$

Proof. Fix $K > 0$. Consider a Poisson arrival process of rate λ on $(-\infty, 0)$ with each arrival having workload with distribution F. Let $\tilde{\Delta}$ denote the total delay experienced by a marked car arriving at time 0 in this system, and let $\tilde{\Delta}^{(K)}$ be the delay of this car when excluding all cars that arrived before time $-K$. Then $\tilde{\Delta}^{(K)}$ has the same distribution as $\Delta^{(K)}$ and $\tilde{\Delta}^{(K)} \uparrow \tilde{\Delta}$ almost surely. It follows that $\lim_{K\to\infty}\mathbb{E}[\Delta^{(K)}] = \mathbb{E}[\tilde{\Delta}]$.

The number of cars N_K arriving in the time interval of length K before the marked arrival has a Poisson distribution with parameter λK. Moreover,

$$\mathbb{P}(\tilde{\Delta}^{(K)} \le s | N_K, W) = \mathbb{P}(W' - KU \le s + W | W)^{N_K}, \quad \text{a.s.} \tag{16}$$

where W' is the workload of a car arriving KU time units before the marked car, and $U \sim U[0,1]$. Conditioning on KU we have that the right hand side of (16) is a.s. equal to

$$\Big(\frac{1}{K}\int_0^K \mathbb{P}(W' \le u + s + W | W)du\Big)^{N_K} = \Big(\frac{1}{K}\int_0^K F(u + s + W)du\Big)^{N_K}.$$

Thus, a.s.

$$\begin{aligned}\mathbb{P}(\tilde{\Delta}^{(K)} \le s | W) &= \sum_{n=0}^\infty \frac{e^{-\lambda K}(\lambda K)^n}{n!}\Big(\frac{1}{K}\int_0^K F(u + s + W)du\Big)^n \\ &= e^{-\lambda K}e^{\lambda\int_0^K F(u+s+W)du} = e^{-\lambda\int_0^K 1 - F(u+s+W)du} \\ &= e^{-\lambda\int_0^K \overline{F}(u+s+W)du}.\end{aligned}$$

By monotone convergence we conclude that a.s.,

$$\begin{aligned}\mathbb{P}(\tilde{\Delta} > s|W) &= 1 - e^{-\lambda \int_0^\infty \overline{F}(u+s+W)du} \\ &= 1 - e^{-\lambda \int_{s+W}^\infty \overline{F}(t)dt}.\end{aligned}$$

Thus, a.s.,

$$\mathbb{E}[\tilde{\Delta}|W] = \int_0^\infty 1 - e^{-\lambda \int_{s+W}^\infty \overline{F}(t)dt} ds = \int_W^\infty 1 - e^{-\lambda \int_v^\infty \overline{F}(t)dt} dv.$$

Taking the expected value of this gives

$$\mathbb{E}[\tilde{\Delta}] = \int_0^\infty \int_w^\infty 1 - e^{-\lambda \int_v^\infty \overline{F}(t)dt} dv dF(w),$$

and a change of order of integration completes the proof. ■

We turn to the proof of Theorem 3. Often these results can be obtained via both coupling and calculus proofs, but we present only one.

Proof of Theorem 3. Let us write

$$f(\lambda) := \mathcal{L}(\lambda, F) = \lambda \int_0^\infty \big(1 - e^{-\lambda H(s)}\big) F(s)\, ds.$$

(a) It is obvious that $f(0) = 0$. Assuming now that $\lambda > 0$, choose $\delta, \epsilon > 0$ such that $0 < F(\delta) < 1$ and $0 < F(\delta + \epsilon) < 1$. Then

$$\mathcal{L}(\lambda, F) \geq \lambda \int_\delta^{\delta+\epsilon} \big(1 - e^{-\lambda H(s)}\big) F(s)\, ds \geq \lambda\epsilon\big(1 - e^{-\lambda H(\delta+\epsilon)}\big) F(\delta) > 0.$$

(b) We first show f is strictly increasing on $[0, \infty)$. Since $f(0) = 0$, and by part (a), $f(h) > 0$ for each $h > 0$, so $f(h) - f(0) > 0$ whenever $h > 0$. Furthermore, for each $\lambda > 0$, and each $h > 0$, a second application of (a) gives

$$f(\lambda + h) - f(\lambda) = \lambda \int_0^\infty (e^{-\lambda H(s)} - e^{-(\lambda+h)H(s)}) F(s) ds + \frac{h f(\lambda + h)}{\lambda + h} \geq \frac{h f(\lambda + h)}{\lambda + h} > 0.$$

We next show f is differentiable on $(0, \infty)$. Fix $\lambda > 0$, and observe that for each $h \neq 0$, and each $s > 0$,

$$\left| \frac{e^{-\lambda H(s)} - e^{-(\lambda+h)H(s)}}{h} \right| \leq H(s)$$

and since H is integrable on $[0, \infty)$ when $\mathbb{E}[W^2] < \infty$, an application of the dominated convergence theorem gives

$$\lim_{h \to 0} \frac{f(\lambda + h) - f(\lambda)}{h} = \int_0^\infty H(s) e^{-\lambda H(s)} F(s) ds + \frac{f(\lambda)}{\lambda}$$

thus proving f is differentiable on $(0, \infty)$.

(c) This follows from a simple coupling argument (it also follows readily from (15)). Generate the arrival process of rate λ and the $\sim F$ workloads of each arrival. Adding α to each workload gives a new system whose workloads have distribution $\sim W + \alpha$, but the delays of individuals are unchanged (delays just start α time units later). This shows that adding a constant to the workloads does not change $\mathcal{L}$.

Returning to the original F system, consider what happens if we speed up (slow down if $\beta < 1$) time by a factor of β. This multiplies the arrival rate by β (and likewise, the time between arrivals by β^{-1}) but also the length of each delay is multiplied by β^{-1}, which implies the loss per unit time is unchanged. This verifies the obvious statement that speeding up (or slowing down) time in the system does not change $\mathcal{L}$.

Now consider what happens if we multiply each workload in the F system by β. Speeding up time by a factor of β gives a process with workloads $\sim F$ again, but now with Poisson arrival rate $\lambda\beta$. By the previous paragraph this speeding up of time does not change $\mathcal{L}$. This shows that scaling the workloads is equivalent to scaling the arrival rate as in the theorem.

(d) This follows from an elementary coupling argument. Compare a single lane system with arrival rate $\lambda_1 + \lambda_2$ to a 2 lane system *with the identical arrival stream* in which each arrival is independently assigned to lane $i \in \{1, 2\}$ with probability $\lambda_i/(\lambda_1 + \lambda_2)$. The delay experienced by any individual in the 2 lane system is less than or equal to the delay that it experiences in the 1 lane system. [1]

(e) On page 945 of [CHR15], the authors show that

$$\mathbb{E}[X_\infty] = \lambda \int_0^\infty (1 + \lambda u)\overline{F}(u)e^{-\lambda H(u)}du$$

and an application of integration by parts reveals that

$$\lambda \int_0^\infty (1 + \lambda u)\overline{F}(u)e^{-\lambda H(u)}du = 1 - e^{-\varrho} + \lambda \int_0^\infty \left(1 - e^{-\lambda H(y)}\right) dy.$$

Furthermore, since it is well-known that $\mathbb{E}[Y_\infty] = \varrho$, we get

$$\mathcal{L}(\lambda, F) = \mathbb{E}[X_\infty] - \mathbb{E}[Y_\infty] = 1 - \varrho - e^{-\varrho} + \lambda \int_0^\infty \left(1 - e^{-\lambda H(y)}\right) dy$$

which in turn means

$$\mathcal{L}(\lambda, F) - \mathcal{L}(\lambda, F') = \varrho' + e^{-\varrho'} - \varrho - e^{-\varrho} + \lambda \int_0^\infty \left(e^{-\lambda H'(y)} - e^{-\lambda H(y)}\right) dy.$$

By the mean value theorem

$$|(\varrho' + e^{-\varrho'}) - (\varrho + e^{-\varrho})| \le |\varrho' - \varrho| \max_{z \in [\varrho, \varrho']} |1 - e^{-z}| \le |\varrho' - \varrho| = \lambda \left|\mathbb{E}[W'] - \mathbb{E}[W]\right|.$$

Similarly

$$\begin{aligned}\left|e^{-\lambda H'(y)} - e^{-\lambda H(y)}\right| &\le \lambda |H'(y) - H(y)| \max_{z \in [H'(y), H(y)]} \left|e^{-\lambda z}\right| \le \lambda |H'(y) - H(y)| \\ &= \lambda \left| \int_y^\infty \left(\overline{F'}(s) - \overline{F}(s)\right) ds \right|.\end{aligned}$$

[1] This can be upgraded to a strict inequality if $\lambda_1, \lambda_2 > 0$.

Finally, by changing the order of integration,

$$\begin{aligned}\big|\mathcal{L}(\lambda,F)-\mathcal{L}(\lambda,F')\big| &\le \lambda|\mathbb{E}[W']-\mathbb{E}[W]| + \lambda^2\int_0^\infty\int_y^\infty \big|\overline{F'}(s)-\overline{F}(s)\big|\,ds\,dy \\ &= \lambda|\mathbb{E}[W']-\mathbb{E}[W]| + \lambda^2\int_0^\infty s\big|\overline{F'}(s)-\overline{F}(s)\big|\,ds. \qquad (17)\end{aligned}$$

This proves the first claim. To prove the second claim, suppose that

$$\int_0^\infty s|\overline{F_n}(s)-\overline{F}(s)|ds\to 0, \quad \text{as } n\to\infty. \qquad (18)$$

Let $\varepsilon>0$ and choose n_0 sufficiently large so that for all $n\ge n_0$,

$$\int_0^\infty s|\overline{F_n}(s)-\overline{F}(s)|ds<\frac{\varepsilon^2}{4}.$$

Then for $n\ge n_0$,

$$\begin{aligned}\big|\mathbb{E}[W_n]-\mathbb{E}[W]\big| &= \Big|\int_0^\infty \overline{F_n}(s)ds-\int_0^\infty\overline{F}(s)ds\Big| \\ &\le \int_0^{\varepsilon/2}|\overline{F_n}(s)-\overline{F}(s)|ds+\int_{\varepsilon/2}^\infty|\overline{F_n}(s)-\overline{F}(s)|ds \\ &\le \frac{\varepsilon}{2}+\int_{\varepsilon/2}^\infty\frac{s}{\varepsilon/2}|\overline{F_n}(s)-\overline{F}(s)|ds \\ &\le \frac{\varepsilon}{2}+\frac{2}{\varepsilon}\int_0^\infty s|\overline{F_n}(s)-\overline{F}(s)|ds\le\varepsilon.\end{aligned}$$

Together with (18) and (17) this verifies that $\mathcal{L}(\lambda,F_n)\to\mathcal{L}(\lambda,F)$ and hence completes the proof.

(f) Assume first that F is the cdf of a random variable taking values $0\le w_1<w_2<\ldots<w_n$. After realizing that the integrand found within the integral is only nonzero in the region $[w_1,w_n)$, and that for each $s\in[w_i,w_{i+1}]$, $H(s)=(w_{i+1}-s)\overline{F}(w_i)+H(w_{i+1})$, we find that

$$\begin{aligned}\mathcal{L}(\lambda,F) &= \lambda\int_{w_1}^{w_n}(1-e^{-\lambda H(s)})F(s)ds \\ &= \lambda\sum_{i=1}^{n-1}F(w_i)\int_{w_i}^{w_{i+1}}(1-e^{-\lambda H(s)})ds \\ &= \lambda\sum_{i=1}^{n-1}F(w_i)\left[(w_{i+1}-w_i)-e^{-\lambda H(w_{i+1})}\int_{w_i}^{w_{i+1}}e^{-\lambda(w_{i+1}-s)\overline{F}(w_i)}ds\right] \\ &= \sum_{i=1}^{n-1}\frac{F(w_i)}{\overline{F}(w_i)}\left[\lambda\overline{F}(w_i)(w_{i+1}-w_i)+e^{-\lambda H(w_i)}-e^{-\lambda H(w_{i+1})}\right].\end{aligned}$$

Combining exponential terms and recalling that $H(w_n)=0$ (so $e^{-\lambda H(w_n)}=1$), we get

$$\begin{aligned}\mathcal{L}(\lambda,F) = &\left[\sum_{i=1}^{n-1}F(w_i)(w_{i+1}-w_i)\right]\lambda-\frac{F(w_{n-1})}{\overline{F}(w_{n-1})} \\ &+\frac{F(w_1)}{\overline{F}(w_1)}e^{-\lambda H(w_1)}+\sum_{i=2}^{n-1}\left[\frac{F(w_i)}{\overline{F}(w_i)}-\frac{F(w_{i-1})}{\overline{F}(w_{i-1})}\right]e^{-\lambda H(w_i)}.\end{aligned}$$

This expression is convex in λ on $[0,\infty)$, because it consists of the sum of an affine function (which is both convex and concave) and $n-1$ convex functions. Note in particular that for this particular type of cdf, $\mathcal{L}(\lambda, F)$ is strictly convex in λ on $[0,\infty)$.

Next, assume F is an arbitrary cdf satisfying $\mathbb{E}[W^2] < \infty$, and approximate W with the sequence $\{W^{(n)}\}_{n\geq 1}$, where for each integer $n \geq 1$,

$$W^{(n)} := n\mathbf{1}_{\{W\geq n\}} + \sum_{k=1}^{n2^n} \frac{k-1}{2^n}\mathbf{1}_{\{(k-1)2^{-n}\leq W<k2^{-n}\}}.$$

Then $W^{(n)} \uparrow W$ pointwise so $\mathbb{E}[(W^{(n)})^k] \to \mathbb{E}[W^k] < \infty$ for $k = 1, 2$ by monotone convergence. Letting F_n represent the cdf of $W^{(n)}$, we clearly have that F_n converges weakly to F as $n \to \infty$, meaning we have pointwise convergence of the CDFs outside of a set of Lebesgue measure zero on $[0,\infty)$. Moreover, since $F_n \geq F_{n+1}$ and $F_n \geq F$ on $[0,\infty)$ for each integer $n \geq 1$, another application of the monotone convergence theorem, gives

$$\lim_{n\to\infty} \int_0^\infty s|\overline{F}_n(s) - \overline{F}(s)|ds = \lim_{n\to\infty} \frac{\mathbb{E}[W^2] - \mathbb{E}[(W^{(n)})^2]}{2} = 0$$

which, by Theorem 3(e), proves $\mathcal{L}(\lambda, F_n) \to \mathcal{L}(\lambda, F)$ as $n \to \infty$. Finally, since the (finite) pointwise limit of a sequence of convex functions must also be convex, we conclude that $\mathcal{L}(\lambda, F)$ is convex in λ on $[0,\infty)$. ■

We next present a comparison result. Let $\mathcal{S}_F$ denote the support of F. For $R \subset \mathbb{R}_+$ we say that a function $g : R \to \mathbb{R}_+$ is an *expansion* if $g(w_2) - g(w_1) \geq w_2 - w_1$ for all $w_2 \geq w_1$ (with $w_2, w_1 \in R$). Note that if $g : R \to \mathbb{R}_+$ is an expansion it must also be increasing.

Lemma 10. *Let $W \sim F$. Let $g : \mathcal{S}_F \to \mathbb{R}_+$ be an expansion, and F_g be the workload cdf of $W' := g(W)$. Then for every $\lambda > 0$, $\mathcal{L}(\lambda, F_g) \geq \mathcal{L}(\lambda, F)$.*

Proof. The proof is by coupling. Suppose that the j-th arrival has delay $\Delta_j \geq 0$. Then

$$\Delta_j = \max_{k\leq j}(T_k + W_k - (T_j + W_j)) = \max_{k\leq j}(W_k - W_j + T_k - T_j).$$

Let Δ'_j denote the corresponding delay of the j-th arrival when we keep the same arrival process, but replace W_k with $W'_k = g(W_k)$ for each $k \in \mathbb{N}$.

If $\Delta_j = 0$ then $\Delta'_j \geq 0 \geq \Delta_j$. Otherwise $\Delta_j > 0$ and there exists some $k < j$ such that $W_k - W_j > T_j - T_k > 0$. For any k with this property, since g is an expansion we have

$$W'_k - W'_j + T_k - T_j = g(W_k) - g(W_j) + T_k - T_j \geq W_k - W_j + T_k - T_j.$$

Hence $\Delta'_j \geq \Delta_j$. The result now follows e.g. by Lemma 9. ■

The next lemma consists of known statements, but we have chosen to include both it and its proof in order to make later arguments easier to follow.

Lemma 11. *Integrals of the function $H(y) = \int_y^\infty \overline{F}(u)\,du$ are related to the first two moments of W and its variance via*

$$\int_0^\infty H(y)\,dy = \frac{1}{2}\mathbb{E}[W^2],\ \int_0^\infty H(s)\overline{F}(s)\,ds = \frac{1}{2}\mathbb{E}[W]^2,\ \int_0^\infty H(y)F(y)\,dy = \frac{1}{2}\mathrm{var}(W).$$

Proof. The first equation follows easily by change of the order of integration in

$$\int_0^\infty H(y)\,dy = \int_0^\infty \int_y^\infty \overline{F}(s)\,ds\,dy = \int_0^\infty s\overline{F}(s)\,ds = \frac{1}{2}\mathbb{E}[W^2].$$

Moreover, letting $D := \int_0^\infty H(s)\overline{F}(s)\,ds$ then

$$\begin{aligned} D &= \int_0^\infty \int_s^\infty \overline{F}(y)\,dy\overline{F}(s)\,ds = \int_0^\infty \int_0^y \overline{F}(s)\,ds\overline{F}(y)\,dy = \int_0^\infty \left(\mathbb{E}[W] - \int_y^\infty \overline{F}(s)\,ds\right)\overline{F}(y)\,dy \\ &= \mathbb{E}[W]\int_0^\infty \overline{F}(y)\,dy - \int_0^\infty \int_y^\infty \overline{F}(s)\,ds\overline{F}(y)\,dy = \mathbb{E}[W]^2 - D, \end{aligned}$$

so the second equality follows. The third is a consequence of the former two, since $\int_0^\infty H(y)F(y)\,dy = \int_0^\infty H(y)\,dy - \int_0^\infty H(y)\overline{F}(y)\,dy$. ■

In what follows we examine the asymptotic behavior of $\mathcal{L}(\lambda, F)$ as $\lambda \to 0$, and as $\lambda \to \infty$, starting with the former one.

Theorem 12. *We have the bound* $\mathcal{L}(\lambda, F) \le \frac{\lambda^2}{2}\mathrm{var}(W)$ *and as* $\lambda \to 0$

$$\mathcal{L}(\lambda, F) \sim \frac{\lambda^2}{2}\mathrm{var}(W). \tag{20}$$

Proof. We first prove the bound. By Lemma 11

$$\frac{\mathcal{L}(\lambda, F)}{\lambda^2} = \underbrace{\int_0^\infty \frac{1 - e^{-\lambda H(s)}}{\lambda} F(s)\,ds}_{B_\lambda} \le \underbrace{\int_0^\infty H(s)F(s)\,ds}_{B} = \frac{1}{2}\mathrm{var}(W).$$

The asymptotic relation follows similarly, since $B_\lambda \to B$ as $\lambda \to 0$ by dominated convergence. ■

We next consider the behavior of $\mathcal{L}(\lambda, F)$ as $\lambda \to \infty$. Let $F^* := \inf\{t \ge 0 : F(t) = 1\}$. In the finite support case ($F^* < \infty$) $\mathcal{L}(\lambda, F)$ increases asymptotically linearly.

Theorem 13. *If* $F^* < \infty$ *then* $\mathcal{L}(\lambda, F) \sim (F^* - \mathbb{E}[W]) \cdot \lambda$ *as* $\lambda \to \infty$.

Proof. By dominated convergence, as $\lambda \to \infty$,

$$\frac{\mathcal{L}(\lambda, F)}{\lambda} = \int_0^{F^*} \left(1 - e^{-\lambda H(s)}\right)F(s)\,ds \to \int_0^{F^*} F(s)\,ds = F^* - \int_0^{F^*} \overline{F}(s)\,ds = F^* - \mathbb{E}[W].$$ ■

We now consider the infinite support case. The integrated tail function $H(y) = \int_y^\infty \overline{F}(s)\,ds$ has $H(0) = \mathbb{E}[W]$ and is continuous and strictly decreasing with $\lim_{y\to\infty} H(y) = 0$. Let $A : [1/\mathbb{E}[W], \infty) \to [0, \infty)$ be the inverse function of the strictly increasing continuous function $1/H(y)$. Then $A(y)$ is strictly increasing and continuous, too, with $A(1/\mathbb{E}[W]) = 0$ and $A(y) \to \infty$ as $y \to \infty$.

We need the concept of slow/regular/rapid variation. A function $g : [0, \infty) \to [0, \infty)$ is *slowly varying* at ∞ if $g(us)/g(u) \to 1$ as $u \to \infty$ for every $s > 0$, and *rapidly varying* at ∞ if $g(us)/g(u) \to 0$ as $u \to \infty$ for every $s > 1$. A function f is *regularly varying* (at ∞) with index $\alpha \in \mathbb{R}$ if $f(u) = u^\alpha g(u)$ for some slowly varying (at ∞) function g.

We write $f \in \mathcal{R}_\alpha$ or $f \in \mathcal{R}_\infty$ if f is regularly varying or rapidly varying respectively. The following result is basically the celebrated Karamata Theorem.

Lemma 14 ([BGT89], Proposition 1.5.10, p.27)**.** *If $\overline{F} \in \mathcal{R}_\alpha$ with $\alpha < -1$ (implying $\mathbb{E}[W^2] < \infty$) then $H(s) \sim \frac{s\overline{F}(s)}{-(1+\alpha)}$ as $s \to \infty$. If $\overline{F}$ is rapidly varying then so is H.*

In the infinite support case, assuming $\overline{F}$ has certain smoothness properties, $\lambda \mapsto \mathcal{L}(\lambda, F)$ increases faster than linear, but is always $o(\lambda^2)$.

Theorem 15. *Suppose that $F^* = \infty$.*

1. *If $\overline{F} \in \mathcal{R}_\alpha$ with $\alpha < -2$ then $A \in \mathcal{R}_{-1/(1+\alpha)}$ and as $\lambda \to \infty$*

$$\mathcal{L}(\lambda, F) \sim \lambda \Gamma\big(\tfrac{\alpha+2}{\alpha+1}\big) A(\lambda). \tag{21}$$

   In particular $\mathcal{L}(\lambda, F) \in \mathcal{R}_{\alpha/(\alpha+1)}$.

2. *If $\overline{F} \in \mathcal{R}_\infty$ then A is slowly varying and as $\lambda \to \infty$,*

$$\mathcal{L}(\lambda, F) \sim \lambda A(\lambda). \tag{22}$$

   In particular $\mathcal{L}(\lambda, F) \in \mathcal{R}_1$.

Proof. 1. Suppose that $\overline{F}$ is regularly varying with index $\alpha < -2$. By Lemma 14 then

$$H(s) \sim \frac{s\overline{F}(s)}{-(\alpha+1)}, \qquad s \to \infty$$

and in particular H is regularly varying with index $\alpha + 1$. For $u \geq 0$ we have

$$\begin{aligned} \mathcal{L}(1/H(u), F) &= \frac{1}{H(u)} \int_0^\infty \big(1 - e^{-\frac{H(r)}{H(u)}}\big) F(r)\, dr \\ &= \frac{u}{H(u)} \int_0^\infty \big(1 - e^{-\frac{H(su)}{H(u)}}\big) F(su)\, ds. \end{aligned} \tag{23}$$

The integrand is bounded by 1 and for any $s > 0$, as $u \to \infty$, $H(su)/H(u) \to s^{\alpha+1}$ and $F(su) \to 1$. Hence

$$\int_0^2 \big(1 - e^{-\frac{H(su)}{H(u)}}\big) F(su)\, ds \to \int_0^2 (1 - e^{-s^{1+\alpha}}) ds.$$

Turning to the tail of the integral, it follows from the Karamata representation theorem that there exists $C > 0$ such that for $\epsilon \in (0, -2-\alpha)$ we can choose u_0 so that for all $u > u_0$ and all $s > 2$

$$\frac{H(su)}{H(u)} \leq C s^{\alpha+1+\epsilon} =: g(s). \tag{24}$$

This is sometimes called a Potter bound. Since $\alpha + 1 + \epsilon < -1$, $\int_2^\infty g(s) ds < \infty$ so by

$$1 - e^{-\frac{H(su)}{H(u)}} \leq \frac{H(su)}{H(u)} \leq g(s),$$

and dominated convergence we get

$$\int_2^\infty \big(1 - e^{-\frac{H(su)}{H(u)}}\big) F(su)\, ds \to \int_2^\infty (1 - e^{-s^{1+\alpha}}) ds.$$

Note that $e^{-s^{1+\alpha}}$ is the cdf of a Frèchet distribution with mean $\Gamma((\alpha+2)/(\alpha+1))$. Altogether we now have

$$\mathcal{L}(1/H(u), F) = \frac{u}{H(u)} \int_0^\infty \left(1 - e^{-\frac{H(su)}{H(u)}}\right) F(su)\, ds \sim \Gamma\left(\tfrac{\alpha+2}{\alpha+1}\right) \frac{u}{H(u)}. \tag{25}$$

Let $\lambda = 1/H(u)$. Since A is the inverse function of $1/H$, $A(\lambda) = u$ and equation (21) follows. Since $1/H$ is regularly varying with index $-(1+\alpha)$ it follows that A is regularly varying with index $-1/(1+\alpha)$ ([BGT89], Theorem 1.5.12), implying that $\mathcal{L}(1/H(u), F)$ is regularly varying with index $1 - 1/(1+\alpha) = \alpha/(1+\alpha)$.

2. For the rapid variation case we again split the integral in (23) up into two integrals,

$$\mathcal{L}(1/H(u), F) = \frac{u}{H(u)} \left(\int_0^1 \left(1 - e^{-\frac{H(su)}{H(u)}}\right) F(su)\, ds + \int_1^\infty \left(1 - e^{-\frac{H(su)}{H(u)}}\right) F(su)\, ds \right).$$

By Lemma 14 the rapid variation of $\overline{F}$ implies rapid variation of H, so $H(su)/H(u) \to \infty$ for $s \in [0,1)$ as $u \to \infty$. Consequently

$$\int_0^1 \left(1 - e^{-\frac{H(su)}{H(u)}}\right) F(su)\, ds \to 1.$$

For the second integral

$$\int_1^\infty \left(1 - e^{-\frac{H(su)}{H(u)}}\right) F(su)\, ds \le \int_1^\infty \frac{H(su)}{H(u)}\, ds = \frac{\int_u^\infty H(s)\, ds}{uH(u)}.$$

By Theorem 1.3.1. in [dH70] the r.h.s. tends to 0 as $u \to \infty$. ■

4 Multiple lanes proofs

Proof of Theorem 6. Let $\boldsymbol{I}$ be a partition, where each I_j is a disjoint union of intervals. Let δ_i be the Lebesgue measure (total length) of I_i, and $c = b - a$. Then the delay is

$$\mathcal{L}(\lambda, F \,|\, \mathbf{I}) = \sum_{i=1}^{\ell} \mathcal{L}(\lambda \delta_i / c, F_{I_i}). \tag{26}$$

Since I_i is a union of intervals whose length is δ_i, F_{I_i} is trivially an expansion of F_{0,δ_i} where $F_{u,v}$ is the cdf of a $U(u,v)$ distribution (the expansion function g simply inserts gaps into the distribution). Applying Lemma 10 to each lane gives

$$\mathcal{L}(\lambda, F \,|\, \mathbf{I}) \ge \sum_{i=1}^{\ell} \mathcal{L}(\lambda \delta_i / c, F_{0,\delta_i}),$$

where (by Theorem 1(c)) the right hand side is equal to the delay rate of a partition routing of F into single intervals of length $\delta_1, \ldots, \delta_\ell$.

Thus we have proved that among partition routings, it is optimal to assign to every lane i a single interval I_i. To prove (a), it remains to show that these should be of equal length. Again, let I_i have length δ_i, and $c = b - a = \sum_{i=1}^{\ell} \delta_i$. By Theorem 1(c), Proposition 4, and the fact that $\frac{1}{\ell} \sum_{i=1}^{\ell} \delta_i^2 \ge \left(\frac{1}{\ell} \sum_{i=1}^{\ell} \delta_i\right)^2 = \frac{c^2}{\ell^2}$ we have

$$\mathcal{L}(\lambda, F_{a,b} \,|\, \mathbf{I}) = \sum_{i=1}^{\ell} \mathcal{L}(\lambda \frac{\delta_i^2 \ell}{c^2}, F_{0,\frac{c}{\ell}}) \ge \sum_{i=1}^{\ell} \mathcal{L}(\lambda \frac{\sum_{i=1}^{\ell} \delta_i^2}{c^2}, F_{0,\frac{c}{\ell}}) \ge \sum_{i=1}^{\ell} \mathcal{L}(\frac{\lambda}{\ell}, F_{0,\frac{c}{\ell}}). \tag{27}$$

By Theorem 1(c) again, the last expression in (27) is the delay obtained by a partition routing of a system with arrival rate λ and workload cdf $F_{a,b}$ into single intervals of equal length. This completes the proof. ■

Proof of Theorem 5. Let the quantiles m_k be such that $F(m_k) = k/\ell$, $k = 0, 1, \ldots, \ell$. The total loss for 'quantile partition routing' (i.e. partition routing with a single interval per lane, with interval endpoints given by the quantiles) is given by

$$\mathcal{L}_1(\lambda, F) = \sum_{k=0}^{\ell-1} \lambda \int_{m_k}^{m_{k+1}} \left(1 - \exp\big(-\lambda \int_s^{m_{k+1}} \Big(\frac{k+1}{\ell} - F(u)\Big)\, du\big)\right) \left(F(s) - \frac{k}{\ell}\right) ds. \tag{28}$$

We compare this with the loss for the optimal random routing (with $q_i = 1/\ell$, see Proposition 4)

$$\begin{aligned}\mathcal{L}_2(\lambda, F) &= \sum_{k=1}^{\ell} \frac{\lambda}{\ell} \int_0^{\infty} \left(1 - \exp\big(-\frac{\lambda}{\ell} \int_s^{\infty} \overline{F}(u)\, du\big)\right) F(s)\, ds. \\ &= \sum_{k=0}^{\ell-1} \lambda \int_{m_k}^{m_{k+1}} \int_0^{\infty} \left(1 - \exp\big(-\frac{\lambda}{\ell} \int_s^{\infty} \overline{F}(u)\, du\big)\right) F(s)\, ds. \end{aligned} \tag{29}$$

Comparing the integrands we will prove that for $m_k \le s \le m_{k+1}$ and $k = 0, 1, 2 \ldots, \ell - 1$

$$\begin{aligned}&\left(1 - \exp\big(-\frac{\lambda}{\ell} \int_s^{m_{k+1}} \big(k + 1 - \ell F(u)\big)\, du\big)\right) (\ell F(s) - k) \\ &\le \left(1 - \exp\big(-\frac{\lambda}{\ell} \int_s^{\infty} (1 - F(u))\, du\big)\right) \ell F(s). \end{aligned} \tag{30}$$

This is true if

$$\int_s^{m_{k+1}} (1 - F(u))\, du + \int_{m_{k+1}}^{\infty} (1 - F(u))\, du \ge \int_s^{m_{k+1}} \big(k + 1 - \ell F(u)\big)\, du$$

i.e. if

$$\int_{m_{k+1}}^{\infty} (1 - F(u))\, du + \int_s^{m_{k+1}} \big((\ell - 1) F(u) - k\big)\, du \ge 0.$$

This inequality certainly holds when s is such that $F(s) \ge \frac{k}{\ell-1}$ (implying $F(u) \ge \frac{k}{\ell-1}$ since $u \ge s$) so the assertion is proved in this case. Hence we assume from now on $F(s) \le \frac{k}{\ell-1}$. Since

$$\int_s^{m_{k+1}} \big(k + 1 - \ell F(u)\big)\, du \le \ell \int_s^{m_{k+1}} \big(1 - F(u)\big)\, du \le \ell \int_s^{\infty} \big(1 - F(u)\big)\, du,$$

it is, considering (30), enough to prove

$$(1 - r)\ell F(s) \ge (1 - r^{\ell})\, (\ell F(s) - k)\,,$$

where $r = \exp\big(-\frac{\lambda}{\ell} \int_s^{\infty} (1 - F(u))\, du\big)$. Equivalently $F(s) \le \frac{k}{\ell} \frac{R}{R-1}$, where $R = 1 + r + r^2 + \ldots + r^{\ell-1}$. Since we assumed $F(u) \le \frac{k}{\ell-1}$, it is sufficient to prove $\frac{k}{\ell-1} \le \frac{k}{\ell} \frac{R}{R-1}$, i.e.

$$\frac{\ell}{\ell - 1} \le \frac{R}{R - 1}$$

which is true because $x \mapsto x/(x-1)$ is monotone decreasing and $\ell \ge R$. ■

Acknowledgements

Part of this work was carried out while MH was a visitor at the Pacific Institute for the Mathematical Sciences (PIMS). BF thanks Clemson University for funding his sabbatical during the Fall 2024 semester, which allowed him to visit MH at PIMS for a week in August 2024, then again for a week in October 2024, in order to work on this project.

References

[AG05] Victor F Araman and Peter W Glynn. Diffusion approximations for the maximum of a perturbed random walk. *Advances in applied probability*, 37(3):663–680, 2005.

[AG06] Victor F Araman and Peter W Glynn. Tail asymptotics for the maximum of perturbed random walk. 2006.

[Asm03] S Asmussen. *Applied Probability and Queues*. Springer-Verlag, 2003.

[BB13] Francois Baccelli and Pierre Brémaud. *Elements of queueing theory: Palm Martingale calculus and stochastic recurrences*, volume 26. Springer Science & Business Media, 2013.

[BGT89] Nicholas H Bingham, Charles M Goldie, and Jef L Teugels. *Regular variation*. Number 27. Cambridge university press, 1989.

[Bri62] L. Brieman. On some probability distributions occurring in traffic flow. *Bull. I.S.I. Paris*, 33:155–161, 1962.

[Bri63] L. Brieman. The poisson tendency in traffic distribution. *Annals of Mathematical Statistics*, 34:308–311, 1963.

[CHR15] Xiuli Chao, Qi-Ming He, and Sheldon Ross. Tollbooth tandem queues with infinite homogeneous servers. *Journal of Applied Probability*, 52(4):941–961, 2015.

[dH70] Laurentius Franciscus Maria de Haan. *On regular variation and its application to the weak convergence of sample extremes*, volume 32. Mathematisch Centrum, 1970.

[Haw66] Alan G. Hawkes. Delay at traffic intersections. *Journal of the Royal Statistical Society, Series B*, 28:202–212, 1966.

[Haw68] Alan G. Hawkes. Gap-acceptance in road traffic. *Journal of Applied Probability*, 5:84–92, 1968.

[Hei94] D. Heidemann. Queue length and delay distributions at traffic signals. *Transportation Research - Part B*, 28B:377–389, 1994.

[JS97] R. Jain and J.M. Smith. Modeling vehicular traffic flow using M/G/C/C state dependent queueing models. *Transportation Science*, 31:324–336, 1997.

[R64] A. Rényi. On two mathematical models of the traffic on a divided highway. *Journal of Applied Probability*, 1:311–320, 1964.

[Ser94] Richard F. Serfozo. Little laws for utility processes and waiting times in queues. *Queueing Systems*, 17:137–181, 1994.

[Ser99] Richard F. Serfozo. *Introduction to Stochastic Networks*. Springer-Verlag, New York, 1999.

[Ser09] Richard Serfozo. *Basics of applied stochastic processes*. Springer Science & Business Media, 2009.

[VWV07] T. Van Woensel and N. Vandaele. Modeling traffic flows with queueing models: a review. *Asia-Pacific Journal of Operational Research*, 24:435–461, 2007.

[Whi91] Ward Whitt. A review of $L = \lambda W$ and extensions. *Queueing Systems*, 9:235–268, 1991.

[Whi92] Ward Whitt. Correction note on $L = \lambda W$. *Queueing Systems*, 12:431–432, 1992.

[Zee] Laplace functional approach to point processes occurring in a traffic model.