The Kingman Bound Basically Holds in the $GI/GI/n$

Amit Harlev, Yige Hong, and Ziv Scully

August 2026

Abstract

We prove simple and explicit bounds on the mean and higher moments of queue length in the first-come, first-served $GI/GI/n$ queue. For all $p \geq 1$, our bound on the pth moment scales as $p^p \cdot \left(\frac{1}{1-\rho}\right)^p$, with the remaining factor depending only on the $(p+1)$th normalized moments of the service time and interarrival time distributions. In particular, our bound does not depend on the number of servers n, so a bound much like the classical Kingman bound for mean queue length in the $GI/GI/1$ also holds in the $GI/GI/n$.

The key new perspective that enables our result is to view (an upper bound on) the queue length process as being parametrized not by a single time dimension, but by $n+1$ time dimensions: one for arrivals, and one for each of the servers. Ordinary passage of time corresponds to advancing all time dimensions at once, but the queue length process is stationary in all dimensions, so we can investigate it by advancing each time dimension individually. This is helpful because shifting just one time dimension gives the queue length process additional monotonicity properties, which lets us bound previously intractable residual interarrival time and residual service time terms.

An initial proof of this result was obtained by GPT-5.6 Sol without human guidance. We digested the initial proof, identified the core new ideas discussed above, and used them to create the much simpler proof herein.

0 Foreword

This document is a "proof note" presenting a bound analogous to the $GI/GI/1$ Kingman bound for the first-come, first-served $GI/GI/n$ queue, uniformly over n and ρ. It is *not intended to be a complete preprint*. In particular, it does not yet contain:

1. A comprehensive discussion of related work.
2. A careful comparison between our techniques and existing techniques in the literature.
3. All of the technical details needed to make every argument fully rigorous.

The first two omissions are especially important because the central proof idea was one-shot by an LLM, specifically GPT-5.6 Sol, without guidance. We therefore regard a careful search and comparison of closely related ideas as even more important here than usual.

We intend to prepare a complete preprint in the near future and to submit it for publication soon afterward. This note serves as an interim public record of the result and its proof while we complete the literature review and comparison with prior techniques. The preprint will also contain a number of extensions to the results presented here; see Section 6.

Why did we feel it necessary to release this note rather than wait until the complete preprint was ready? There are two reasons.

The first is practical. Amit is writing a blog post about AI-generated proofs and the questions they raise for the mathematical community, particularly the community surrounding the SIGMETRICS conference (the research community in which he is most involved). His experience prompting GPT-5.6 Sol for this result and then working with Yige and Ziv to verify, refine, and explain it will serve as a case study in that post. A public document was needed for the post to reference.

The second reason relates to priority and credit. Because an LLM can one-shot this result, another person could potentially publish the same result very quickly with a minimally checked or poorly written LLM-generated proof. We have collectively spent over one hundred hours understanding and verifying the argument, introducing new mathematical objects that clarify it and separate the existing ideas from the new contribution, and refining the exposition. Although the key trick remained unchanged, the proof and presentation differ greatly from the original LLM-generated version. We would be disappointed if our work received little or no recognition just because someone else released a rough version first.

This exposes a troubling tension. Auditing an LLM-generated proof for correctness, understanding its relationship to the literature, and writing it up well all require a significant time commitment. However, releasing the raw output quickly is the safer strategy for establishing priority, and thus receiving credit, as the mathematical community currently places considerable weight on who first releases a result when assigning credit. One of the goals of the blog post is to discuss the problematic incentive this creates.

Lastly, although this note does not yet contain a comprehensive discussion of related work, we want to emphasize the close relationship between our proof and that of Hong [3]. Despite differences in presentation, our proof can be viewed as that argument augmented by one new idea. This idea replaces the worst-case residual-service bound with a new bound, allowing us to assume only finite second moments in place of a bounded worst-case mean residual time.

1 Introduction

Consider a first-come, first-served $GI/GI/n$ queue with n homogeneous servers. Jobs arrive according to a renewal process whose interarrival times are distributed as A, and their service times (a.k.a. sizes) are independent copies of a random variable S. Define

$$\lambda \coloneqq \frac{1}{\mathbf{E}[A]}, \qquad \mu \coloneqq \frac{1}{\mathbf{E}[S]}, \qquad \rho \coloneqq \frac{\lambda}{n\mu}.$$

Thus, λ is the arrival rate, each server works at rate μ, and ρ is the load. We assume throughout that $\rho < 1$. Let L denote the stationary number of jobs waiting for service in the standard $GI/GI/n$ queue.

Kingman [5] proved that in the single-server setting where $\rho = \lambda/\mu$,

$$\mathbf{E}[L_{GI/GI/1}] \le \frac{\rho^2 \mathbf{E}[(\mu S)^2] + \mathbf{E}[(\lambda A)^2] - (1+\rho^2)}{2(1-\rho)}.$$

If furthermore the arrival process is Poisson, the Pollaczek–Khinchine [4, 8, 9] formula gives the exact mean

$$\mathbf{E}[L_{M/G/1}] = \frac{\rho^2 \mathbf{E}[(\mu S)^2]}{2(1-\rho)}.$$

In this note, we prove analogous bounds on $\mathbf{E}[L]$ for the multiserver setting. Define

$$C_{\text{service}} \coloneqq \sup_{t>0} \frac{\mathbf{E}\left[\int_0^{S\wedge t} (S-a)\,\mathrm{d}a\right]}{\mathbf{E}[S \wedge t]},$$

where $x \wedge y \coloneqq \min(x, y)$. One should think of C_{service} as the largest *cutoff-averaged residual service time*. For a fixed cutoff t, the ratio averages the residual service time $S - a$ over all elapsed times a in $(0, S \wedge t]$. The supremum then selects the cutoff producing the largest such average.

Theorem 1.1. *Suppose that $\rho < 1$ and*

$$\mathbf{E}[S^2] + \mathbf{E}[A^2] < \infty.$$

Then

$$\mathbf{E}[L] \le \frac{\rho\mu C_{\text{service}} + \frac{1}{2}\mathbf{E}[(\lambda A)^2] - \rho}{1-\rho} \le \frac{\rho \mathbf{E}[(\mu S)^2] + \frac{1}{2}\mathbf{E}[(\lambda A)^2] - \rho}{1-\rho}.$$

Furthermore, if the arrival process is Poisson,

$$\mathbf{E}[L] \le \frac{\rho}{1-\rho}\mu C_{\text{service}} \le \frac{\rho}{1-\rho}\mathbf{E}[(\mu S)^2].$$

Notably, despite holding uniformly over the number of servers n, our bounds remain remarkably close to the classical single-server bounds. The comparison becomes even tighter when the service-time distribution has *increasing mean residual life* (IMRL). This means that

$$a \mapsto \mathbf{E}[S - a \mid S > a]$$

is non-decreasing over the values of a for which $\mathbf{P}[S > a] > 0$. In other words, the expected remaining service time does not decrease as the amount of service already received increases. If S is IMRL, C_{service} equals the mean equilibrium residual service time.

Corollary 1.2. *If S is IMRL, the bound in Theorem 1.1 becomes*

$$\mathbf{E}[L] \le \frac{\rho \mathbf{E}[(\mu S)^2] + \mathbf{E}[(\lambda A)^2] - 2\rho}{2(1-\rho)}.$$

If, in addition, the arrival process is Poisson, then

$$\mathbf{E}[L] \le \frac{\rho}{2(1-\rho)}\mathbf{E}[(\mu S)^2].$$

Our proof also provides analogous bounds on higher moments of L.

Theorem 1.3. *Fix $p \geq 1$ and suppose that*

$$\mathbf{E}[(\mu S)^{p+1}] + \mathbf{E}[(\lambda A)^{p+1}] < \infty.$$

Then

$$\mathbf{E}[L^p] \leq \left[\frac{p}{1-\rho}\left(\rho \mathbf{E}[(\mu S)^{p+1}]^{1/p} + \mathbf{E}[(\lambda A)^{p+1}]^{1/p}\right)\right]^p.$$

1.1 A very brief summary of past results

The search for a simple, universal multiserver analogue of Kingman's bound has a history spanning more than fifty years. As this document is intended as a "proof note" rather than a complete paper, we keep this summary intentionally brief. For a substantially more complete history of the problem and the many results preceding these uniform bounds, we refer the interested reader to the introduction of Li and Goldberg [6].

The first simple and explicit bound with $1/(1-\rho)$ scaling that holds uniformly over both n and ρ was proved by Li and Goldberg [6]. Fix $\varepsilon \in (0, 1/2)$ and suppose that

$$\mathbf{E}[A^2] < \infty \qquad \text{and} \qquad \mathbf{E}[S^{2+\varepsilon}] < \infty.$$

Under the standing assumptions that $\rho < 1$ and the steady-state queue length exists, they obtain

$$\mathbf{E}[L] \leq \frac{1}{1-\rho}\left[2.1 \times 10^{21} \mathbf{E}[(\mu S)^2]\left((\mathbf{E}[(\mu S)^2])^{1+\varepsilon} + \mathbf{E}[(\mu S)^{2+\varepsilon}]\right)\varepsilon^{-4} + 49\mathbf{E}[(\lambda A)^2]\right].$$

Thus, their mean bound depends on the second moment of the interarrival time and a $(2+\varepsilon)$th moment of the service time. They also establish polynomial tail bounds and corresponding higher-moment bounds; the order of the queue-length moment they can control increases as additional service-time moments are assumed finite. The principal drawback is the very large numerical constant.

More recently, Hong [3] obtained a substantially smaller and more interpretable bound under stronger assumptions on the service-time distribution. Define

$$R_s^{\max} \coloneqq \sup_{t \geq 0} \mathbf{E}[\mu S - t \mid \mu S \geq t] \quad \text{and} \quad R_a^{\min} \coloneqq \inf_{t \geq 0} \mathbf{E}[\lambda A - t \mid \lambda A \geq t].$$

Hong [3] assumes that

$$\mathbf{E}[A^2] < \infty, \qquad R_s^{\max} < \infty,$$

and that S is non-lattice. Under these assumptions, a consequence of his main theorem is

$$\mathbf{E}[L] \leq \frac{R_s^{\max} + \frac{1}{2}\mathbf{E}[(\lambda A)^2] - \rho}{1-\rho} - R_a^{\min}.$$

For Poisson arrivals, this simplifies to

$$\mathbf{E}[L] \le \frac{R_s^{\max}}{1-\rho}.$$

Although the bound of Hong [3] has much smaller constants than that of Li and Goldberg [6], the condition $R_s^{\max} < \infty$ is considerably stronger than the existence of any fixed service-time moment. In fact, this condition implies that S has an exponentially decaying tail and hence moments of every order [3, Appendix A].

Like the bound of Hong [3], our mean bound has small explicit constants, but it controls the service contribution through the cutoff-averaged quantity μC_{service} rather than the worst-case mean residual time $R_s^{\max}$. This allows us to assume only finite second moments of A and S. Our higher-moment result similarly bounds $\mathbf{E}[L^p]$ using only the $(p+1)$th moments of the two distributions.

2 The modified $GI/GI/n$ queue and virtual queues

Rather than study the $GI/GI/n$ queue directly, we study a modified $GI/GI/n$ queue with a simpler stochastic structure. The construction is standard in the queueing literature and was used first in Li and Goldberg [6] and later in Hong [3] to establish $1/(1-\rho)$ scaling.

Crucially, analyzing the modified system is sufficient for obtaining bounds on the standard queue. The stationary queue length in the modified $GI/GI/n$ system stochastically dominates that of the standard $GI/GI/n$ queue; see Gamarnik and Goldberg [2, Proposition 1]. We therefore work exclusively with the modified system throughout the remainder of the paper. Any upper bound proved for its queue length immediately yields the same bound for the standard $GI/GI/n$ queue.

In the modified $GI/GI/n$ queue, at any time a server would become idle, it instead "goes on vacation" for an interval with length independently sampled from S (the job size distribution). Vacations are not preemptible: servers will remain on vacation until the end of the interval even if a new job arrives into the system. Thus, in the modified $GI/GI/n$, each server's sequence of service-completion and vacation-ending epochs forms a renewal process with i.i.d. increments distributed as S.

For purposes of tracking queue length, these epochs admit the following equivalent interpretation. Each server generates a renewal process of completion tokens with interrenewal distribution S. We will refer to these as token streams. When a completion token is produced, it instantaneously completes one job in the queue and removes that job from the system. If the queue is empty, the token is wasted and has no effect. The n token streams are mutually independent, and the arrival epochs form a separate renewal process with interrenewal distribution A, independent of every token stream. We refer collectively to the arrival process and the n token streams as the *primitive processes*, since they are the basic stochastic inputs of the model.

To represent these processes, it is useful to work on a canonical sample space that

records all of their event times. A sample point is written as

$$\mathbf{x} = (\vec{x}_a, \vec{x}_1, \dots, \vec{x}_n),$$

where $\vec{x}_a$ records the arrival process and $\vec{x}_\ell$ records the token stream of server ℓ. Together, these vectors specify the entire past and future of every primitive process and therefore determine the complete queue-length sample path.

More explicitly, write

$$\vec{x}_a = (\dots, U^{-1}, U^0, U^1, \dots)$$

for the vector of arrival times and

$$\vec{x}_\ell = (\dots, T_\ell^{-1}, T_\ell^0, T_\ell^1, \dots), \qquad \ell = 1, \dots, n,$$

for the vectors of token times. We index these vectors so that

$$U^0 \le 0 < U^1, \qquad T_\ell^0 \le 0 < T_\ell^1.$$

Thus, the event with index zero is the most recent event at or before time zero, while the event with index one is the next event after time zero.

We next define the stationary probability law on this sample space. Let $\mathbf{P}_{a,\mathrm{stat}}[\cdot]$ denote the stationary renewal law of the arrival vector $\vec{x}_a$, and, for each server ℓ, let $\mathbf{P}_{\ell,\mathrm{stat}}[\cdot]$ denote the stationary renewal law of the token vector $\vec{x}_\ell$. Because the primitive processes are mutually independent, their joint stationary law is

$$\mathbf{P}_\pi[\cdot] \coloneqq \mathbf{P}_{a,\mathrm{stat}}[\cdot] \otimes \bigotimes_{\ell=1}^{n} \mathbf{P}_{\ell,\mathrm{stat}}[\cdot].$$

Here $\otimes$ denotes the product of the displayed probability measures: under $\mathbf{P}_\pi[\cdot]$, the arrival vector and token vectors are sampled independently according to their respective stationary laws. We write $\mathbf{E}_\pi[\cdot]$ for expectation under $\mathbf{P}_\pi[\cdot]$. Throughout the paper, unless explicitly stated otherwise, stationarity refers to stationarity under $\mathbf{P}_\pi[\cdot]$.

For $s < t$, define

$$\mathrm{arr}(s, t; \mathbf{x}) \coloneqq \text{number of arrivals in } [s, t)$$

and, for each server ℓ,

$$\mathrm{tok}_\ell(s, t; \mathbf{x}) \coloneqq \text{number of tokens from token stream } \ell \text{ in } [s, t).$$

Note that $\mathrm{arr}(s, t; \mathbf{x})$ depends only on $\vec{x}_a$, while $\mathrm{tok}_\ell(s, t; \mathbf{x})$ depends only on $\vec{x}_\ell$. Define the queue-length functional Q by

$$Q(\mathbf{x}) \coloneqq \sup_{s<0} \left[\mathrm{arr}(s, 0; \mathbf{x}) - \sum_{\ell=1}^{n} \mathrm{tok}_\ell(s, 0; \mathbf{x}) \right]. \tag{2.1}$$

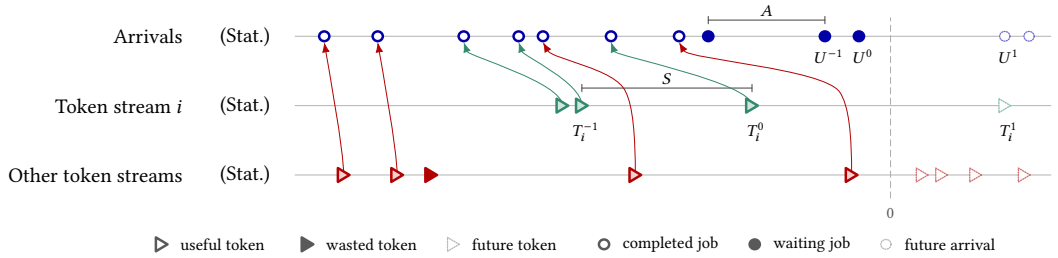


Figure 2.1. A stationary realization of the modified queue. The labels (Stat.) indicate that each primitive process is sampled from its stationary renewal law. i denotes an arbitrary representative server. The arrival epochs U^{-1}, U^0, U^1 and the token epochs T_i^{-1}, T_i^0, T_i^1 illustrate the indexing of the coordinates of $\vec{x}_a$ and $\vec{x}_i$, respectively: U^0 and T_i^0 are the most recent corresponding events at or before time zero, while U^1 and T_i^1 are the next events after time zero. The arrows display the matching between useful tokens and completed jobs, while the labeled intervals illustrate that consecutive arrival and token epochs are separated by increments distributed as A and S, respectively.

The supremum accounts for the fact that tokens cannot be stored: tokens produced before the queue most recently became nonempty may have been wasted and are therefore irrelevant. The queue length is consequently the largest excess of arrivals over tokens on an interval ending at time zero. Because the counting intervals exclude their right endpoints, events at time zero are not included in these counts. Thus, $Q(\mathbf{x})$ is the queue length immediately before time zero.

We now use $Q(\mathbf{x})$ to define the queue process. We will also use $Q(\mathbf{x})$ to define two other processes, which we refer to as virtual queues and will be essential to our analysis. For $t \in \mathbb{R}$, let $\mathrm{adv}(\mathbf{x}, t)$ denote the sample point obtained by advancing every primitive process forward by t units of time. That is, $\mathrm{adv}(\mathbf{x}, t)$ is obtained by replacing every event time u appearing in $\mathbf{x}$ with $u - t$. For example, token stream ℓ becomes

$$(\ldots, T_\ell^{-1} - t, T_\ell^0 - t, T_\ell^1 - t, \ldots).$$

Define the queue process by

$$Q(t; \mathbf{x}) \coloneqq Q(\mathrm{adv}(\mathbf{x}, t)).$$

When the sample point $\mathbf{x}$ is clear from context, we abbreviate $Q(t; \mathbf{x})$ to $Q(t)$. We will also use $Q \coloneqq Q(0)$ for convenience. Because Q gives the queue length immediately before time zero, $Q(t)$ gives the queue length immediately before time t and is therefore left-continuous.

Figure 2.1 gives a concrete representation of the queue at time zero. A job contributes to the queue precisely when it has not been matched to a token by the dashed time-zero line. There are three such jobs in the realization shown, and hence $Q = 3$. A token produced when no job is waiting remains unmatched and is wasted, so it cannot offset a later arrival.

For a fixed server i, let $\mathrm{adv}_i(\mathbf{x}, t)$ denote the sample point obtained by advancing only token stream i by t, while holding the arrival process and all other token streams fixed.

Similarly, let $\mathrm{adv}_a(\mathbf{x}, t)$ denote the sample point obtained by advancing only the arrival process by t. We define the corresponding virtual queue processes by

$$Q_i(t; \mathbf{x}) \coloneqq Q(\mathrm{adv}_i(\mathbf{x}, t)) \quad \text{and} \quad Q_a(t; \mathbf{x}) \coloneqq Q(\mathrm{adv}_a(\mathbf{x}, t)),$$

with the former referred to as the virtual token queue and the latter as the virtual arrival queue. When $\mathbf{x}$ is clear from context, we abbreviate these to $Q_i(t)$ and $Q_a(t)$. Here t represents virtual time: only the indicated primitive process advances, while every other primitive process remains fixed. Just like Q, these processes are left-continuous. We will also use $Q_i \coloneqq Q_i(0)$ and $Q_a \coloneqq Q_a(0)$.

Each virtual queue behaves like a modified single-server queue in virtual time. To see why, fix the arrival process and all token streams other than token stream i, and advance only token stream i. By the definition of Q,

$$Q_i(t; \mathbf{x}) = \sup_{s<0} \left[\mathrm{arr}(s, 0; \mathbf{x}) - \sum_{\ell \neq i} \mathrm{tok}_\ell(s, 0; \mathbf{x}) - \mathrm{tok}_i(s, 0; \mathrm{adv}_i(\mathbf{x}, t)) \right].$$

This representation makes the dynamics of $Q_i(t)$ clear both at epochs of token stream i and between consecutive tokens:

- At each token epoch in virtual time, one token of token stream i passes through time zero. It removes one job if $Q_i(t)$ is nonempty and is wasted otherwise.
- Between consecutive tokens, no new token crosses time zero. Thus, for every $s < 0$, advancing token stream i can only decrease $\mathrm{tok}_i(s, 0; \mathrm{adv}_i(\mathbf{x}, t))$. Every other term inside the supremum remains fixed, so $Q_i(t)$ can increase but not decrease between consecutive tokens.

We call these upward jumps in the virtual queue length *virtual arrivals*. Viewed in virtual time, $Q_i(t)$ is therefore a modified single-server queue whose service tokens are supplied by token stream i and whose arrivals are the virtual arrivals. Analogously, $Q_a(t)$ increases when an arrival crosses time zero and can only decrease between consecutive arrivals, at epochs that we call *virtual departures*.

Comparing Figures 2.1 and 2.2 shows how a virtual arrival is created. As token stream i advances from virtual time 0 to virtual time 1.3, the matching arrows change: the affected tokens are reassigned among the arrivals, and the indicated token eventually occurs too early to remove any job and is wasted. This change propagates through the matching until the job at U^{-4} is left unmatched, causing $Q_i(t)$ to increase by one. We retain the hollow marker at U^{-4} to display its completed status in the original matching and represent its reappearance in the virtual queue by the yellow virtual arrival on the token-stream-i line.

At virtual time zero, both virtual queues agree pathwise with the real queue:

$$Q(0; \mathbf{x}) = Q_i(0; \mathbf{x}) = Q_a(0; \mathbf{x}) = Q(\mathbf{x}).$$

Now suppose that $\mathbf{x}$ is sampled under $\mathbf{P}_\pi[\cdot]$. Each primitive process is then sampled independently from its stationary renewal law. Advancing any one primitive process, or advancing all of them simultaneously, therefore leaves their joint distribution unchanged.

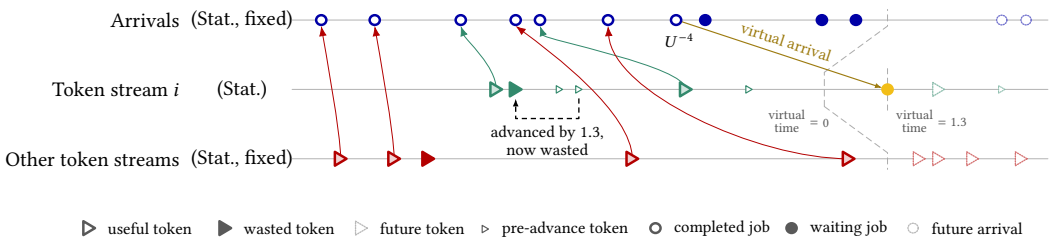


Figure 2.2. The stationary realization from Figure 2.1 after advancing only token stream i by 1.3 units of virtual time. The arrival process and all other token streams remain fixed exactly as in Figure 2.1. The dashed diagonal guides account for the different clocks: time zero remains fixed on the arrival and other-token rows, while the virtual-time origin on token stream i moves with that token stream. Consequently, the fixed processes at time zero align with token stream i at virtual time 1.3. The small green triangles mark the token locations in Figure 2.1, and the yellow circle records the resulting virtual arrival.

Consequently, the preceding pathwise equality at time zero extends to an equality in distribution at every real or virtual time.

Lemma 2.1. *Under any probability law on the primitive processes,*

$$Q = Q_i = Q_a.$$

Under $\mathbf{P}_\pi[\cdot]$, the processes $Q(t)$, $Q_i(t)$, and $Q_a(t)$ are stationary in their respective time parameters, and so the above implies

$$Q(t) \overset{\mathrm{d}}{=} Q_i(t) \overset{\mathrm{d}}{=} Q_a(t) \overset{\mathrm{d}}{=} Q,$$

for every $t \in \mathbb{R}$. In particular, for every $p > 0$ for which the corresponding moment is finite,

$$\mathbf{E}_\pi[Q(t)^p] = \mathbf{E}_\pi[Q_i(t)^p] = \mathbf{E}_\pi[Q_a(t)^p] = \mathbf{E}_\pi[Q^p].$$

Lemma 2.1 captures the central reason for introducing the virtual queues: we can recover moments of the real queue by analyzing the simpler dynamics of either virtual queue. We will do this using rate-conservation arguments applied to the virtual queues and then transfer the resulting identities to the real queue. These arguments require the distribution of the system as seen at a typical token, arrival, virtual arrival, or virtual departure, so we next introduce the corresponding Palm measures.

We begin with the Palm measures associated with the primitive renewal processes. Let $\mathbf{P}_{a,\mathrm{Palm}}[\cdot]$ denote the Palm law of the arrival vector $\vec{x}_a$. Under this law, time zero is an arrival epoch, so $U^0 = 0$. Because the arrival process is independent of the token streams, the joint law of the complete system as seen at a typical arrival is

$$\mathbf{P}_a[\cdot] \coloneqq \mathbf{P}_{a,\mathrm{Palm}}[\cdot] \otimes \bigotimes_{\ell=1}^{n} \mathbf{P}_{\ell,\mathrm{stat}}[\cdot].$$

We write $\mathbf{E}_a[\cdot]$ for expectation under $\mathbf{P}_a[\cdot]$.

Similarly, let $\mathbf{P}_{i,\mathrm{Palm}}[\cdot]$ denote the Palm law of token stream i. Under this law, time zero is a token epoch of server i, so $T_i^0 = 0$. The joint law of the complete system as seen at a typical token of server i is

$$\mathbf{P}_i[\cdot] \coloneqq \mathbf{P}_{a,\mathrm{stat}}[\cdot] \otimes \mathbf{P}_{i,\mathrm{Palm}}[\cdot] \otimes \bigotimes_{\ell \neq i} \mathbf{P}_{\ell,\mathrm{stat}}[\cdot].$$

We write $\mathbf{E}_i[\cdot]$ for expectation under $\mathbf{P}_i[\cdot]$. Since the queue processes are left-continuous, Q, Q_i, and Q_a under these Palm measures denote their values immediately before the event at time zero; we use $Q(0+)$, $Q_i(0+)$, and $Q_a(0+)$ for the corresponding post-event values.

We now introduce the Palm measures associated with virtual arrivals and virtual departures. To avoid undue notation, we assume that no two virtual arrivals or departures occur at the same virtual time.[1]

Because $Q_i(t)$ is stationary in virtual time, its virtual arrivals form a stationary point process. Let $\lambda_{i\uparrow}$ denote the rate of this process, and let $\mathbf{P}_{i\uparrow}[\cdot]$ and $\mathbf{E}_{i\uparrow}[\cdot]$ denote its Palm probability and expectation. Thus, under $\mathbf{P}_{i\uparrow}[\cdot]$, time zero is a typical virtual arrival and

$$Q_i(0+) = Q_i + 1.$$

Similarly, the virtual departures of the stationary process $Q_a(t)$ form a stationary point process. Let $\lambda_{a\downarrow}$ denote its rate, and let $\mathbf{P}_{a\downarrow}[\cdot]$ and $\mathbf{E}_{a\downarrow}[\cdot]$ denote its Palm probability and expectation. Thus, under $\mathbf{P}_{a\downarrow}[\cdot]$, time zero is a typical virtual departure and

$$Q_a(0+) = Q_a - 1.$$

Unlike the Palm laws of the primitive renewal processes, $\mathbf{P}_{i\uparrow}[\cdot]$ and $\mathbf{P}_{a\downarrow}[\cdot]$ generally do not factor into independent marginal laws, because the corresponding virtual events depend on the complete collection of primitive processes.

We now compute the rate of virtual arrivals and departures, which we will need in the next section to use rate conservation.

Lemma 2.2. *The virtual-arrival and virtual-departure rates satisfy*

$$\lambda_{i\uparrow} = \frac{\lambda}{n} = \mu\rho, \qquad \lambda_{a\downarrow} = \lambda.$$

Proof. In stationarity, the aggregate rate at which useful tokens remove jobs from the real queue $Q(t)$ equals the arrival rate λ, since the queue has zero long-run drift. By symmetry,

[1] This assumption holds, for example, under standard continuous interarrival and service distributions. It can fail for lattice distributions, and more generally whenever several queue changes occur at the same virtual time. If the virtual queue increases from q to $q + r$ at a single virtual time, one may treat the change as r simultaneous virtual arrivals, associated with the successive levels $q + 1, \ldots, q + r$; downward changes can be treated analogously as simultaneous virtual departures. The Palm and rate-conservation arguments below then extend without substantive change, but require additional notation to index these intermediate states between the pre-event state at time 0− and the post-event state at time 0+.

useful tokens from token stream i therefore occur at rate λ/n. Equivalently,

$$\mu \mathbf{P}_i[Q > 0] = \frac{\lambda}{n}. \tag{2.2}$$

By Lemma 2.1, $\mathbf{P}_i[Q > 0] = \mathbf{P}_i[Q_i > 0]$. Since token stream i also has rate μ in virtual time, the downward jumps of $Q_i(t)$ also occur at rate λ/n. Thus,

$$\lambda_{i\uparrow} = \frac{\lambda}{n} = \mu\rho.$$

Similarly, arrival epochs occur at rate λ and produce the upward jumps of $Q_a(t)$. Since $Q_a(t)$ is stationary in virtual time, its downward jump rate equals its upward jump rate. Therefore, $\lambda_{a\downarrow} = \lambda$. □

With these Palm measures and rates in hand, we are now ready to bound $\mathbf{E}_\pi[Q^p]$ using rate-conservation arguments for the two virtual queues.

3 Bounding $\mathbf{E}_\pi[Q^p]$ with rate conservation

Technical note: for a rigorous proof, one must carefully apply truncation arguments throughout the rest of this note. The resulting bounds are uniform in the truncation levels, which may then be sent to infinity using monotone convergence, dominated convergence and other standard limiting arguments. We suppress such truncations from the argument to keep the proof focused on the main ideas.

Our first step will be to use arrival-departure symmetry in each of $Q(t)$, $Q_i(t)$, and $Q_a(t)$ to relate their pth moments under the different Palm measures.

Lemma 3.1. *For every $p \geq 1$,*

$$\rho\mathbf{E}_{i\uparrow}[(Q_i + 1)^p] = \mathbf{E}_i[Q^p] = \rho\mathbf{E}_a[(Q + 1)^p] = \rho\mathbf{E}_{a\downarrow}[Q_a^p].$$

Note that by Lemma 2.1, any occurrence of Q, Q_i and Q_a in the above equality can be replaced by either of the other two.

Proof. Since $Q_i(t)$ is stationary, its upward and downward crossings must balance. Weighting each crossing by the pth power of the crossed value, the average contribution at virtual arrivals must equal the average contribution at tokens of token stream i. Therefore,

$$\lambda_{i\uparrow}\mathbf{E}_{i\uparrow}[(Q_i + 1)^p] = \mu\mathbf{E}_i[Q_i^p].$$

At virtual time zero, $Q_i = Q$, so $\mathbf{E}_i[Q_i^p] = \mathbf{E}_i[Q^p]$. Using $\lambda_{i\uparrow} = \mu\rho$ from Lemma 2.2, we obtain

$$\rho\mathbf{E}_{i\uparrow}[(Q_i + 1)^p] = \mathbf{E}_i[Q^p].$$

Similarly, stationarity of $Q(t)$ implies that the average contribution at arrivals must balance the aggregate average contribution at tokens from all n token streams. Hence,

$$\lambda \mathbf{E}_a[(Q+1)^p] = n\mu \mathbf{E}_i[Q^p].$$

Since $\lambda = n\mu\rho$, it follows that

$$\rho \mathbf{E}_a[(Q+1)^p] = \mathbf{E}_i[Q^p].$$

Finally, stationarity of $Q_a(t)$ implies that the average contribution at arrivals must balance the average contribution at virtual departures. Therefore,

$$\lambda \mathbf{E}_a[(Q_a+1)^p] = \lambda_{a\downarrow} \mathbf{E}_{a\downarrow}[Q_a^p].$$

At virtual time zero, $Q_a = Q$, and $\lambda_{a\downarrow} = \lambda$ by Lemma 2.2. Thus,

$$\mathbf{E}_a[(Q+1)^p] = \mathbf{E}_{a\downarrow}[Q_a^p]. \qquad \square$$

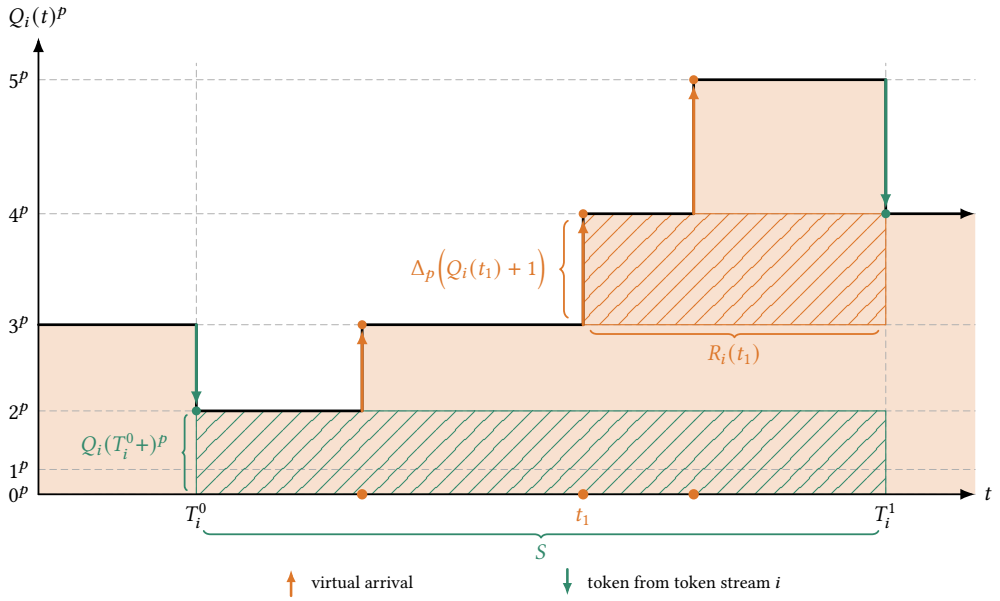


Figure 3.1. The process $Q_i(t)^p$ over one renewal cycle of token stream i, from T_i^0 to T_i^1. Upward orange arrows mark virtual arrivals, while downward green arrows mark tokens from token stream i. The green hatched rectangle has area $SQ_i(T_i^0+)^p$. The orange hatched rectangle has area $R_i(t_1)\Delta_p\big(Q_i(t_1)+1\big)$, where its width is the residual time until the next token and its height is the increase in Q_i^p caused by the virtual arrival at t_1.

We can now use the virtual queues to bound $\mathbf{E}_\pi[Q^p]$. Since $Q_i(t)$ is stationary, the Palm inversion formula (which is a generalization of the renewal-reward theorem) expresses

its pth moment as the expected area under $Q_i(t)^p$ during a single renewal cycle of token stream i, divided by the expected cycle length:

$$\mathbf{E}_\pi[Q_i^p] = \frac{\mathbf{E}_i\left[\text{area under } Q_i(t)^p \text{ on } [T_i^0, T_i^1]\right]}{\mathbf{E}[S]}. \tag{3.1}$$

Our plan is to partition this area into contributions associated with tokens and virtual arrivals. To do so, we will need notation for the residual time until the next token. For $t \in \mathbb{R}$, let $R_i(t)$ denote the residual virtual time from t until the next token of token stream i. At a token epoch, we take $R_i(t) = 0$, so that $R_i(t)$ is left-continuous. We also use $R_i \coloneqq R_i(0)$.

As illustrated in Figure 3.1, the area under $Q_i(t)^p$ during the renewal interval $[T_i^0, T_i^1]$ can be partitioned into two types of rectangles:

1. At the initial token time T_i^0, a base rectangle extending across the entire renewal interval, with area $SQ_i(T_i^0+)^p$.
2. For each virtual arrival time $t \in (T_i^0, T_i^1)$, an additional rectangle extending from t until T_i^1, with area $R_i(t)\Delta_p\Big(Q_i(t) + 1\Big)$, where $\Delta_p(q) \coloneqq q^p - (q-1)^p$.

The first kind of rectangle contributes expected area

$$\mathbf{E}_i[SQ_i(0+)^p] = \mathbf{E}[S]\mathbf{E}_i[Q_i(0+)^p].$$

The second kind of rectangle has expected area

$$\mathbf{E}_{i\uparrow}\left[R_i\Delta_p\big(Q_i + 1\big)\right]$$

and such rectangles occur at rate $\lambda_{i\uparrow} = \lambda/n$ (Lemma 2.2) over the interval. Thus, they contribute expected total area

$$\mathbf{E}[\#\text{ of rectangles}] \times \mathbf{E}[\text{area of rectangles}] = \left(\frac{\lambda}{n}\mathbf{E}[S]\right) \cdot \mathbf{E}_{i\uparrow}\left[R_i\Delta_p\big(Q_i + 1\big)\right].$$

Adding the area contributions from both kinds of rectangles, plugging into (3.1), and using Lemma 2.1 yields the following lemma.[2]

Lemma 3.2. *For $p \geq 1$, define $\Delta_p(q) \coloneqq q^p - (q-1)^p$ for $q \geq 1$. Then*

$$\mathbf{E}_\pi[Q^p] = \mathbf{E}_i[Q_i(0+)^p] + \frac{\lambda}{n}\mathbf{E}_{i\uparrow}\Big[R_i \cdot \Delta_p\big(Q_i + 1\big)\Big].$$

Similarly, $Q_a(t)$ is stationary, so we can compute $\mathbf{E}_\pi[Q_a^p]$ by reasoning about areas. Define $R_a(t)$ to be the residual virtual time from t until the next arrival, taking $R_a(t) = 0$ at an arrival epoch so that the process is left-continuous. We also use $R_a \coloneqq R_a(0)$. As illustrated in Figure 3.2, the area under $Q_a(t)^p$ during the renewal interval $[U^0, U^1]$ can be obtained by first placing a base rectangle at U^0 extending over the entire interval and then removing a rectangular cutout for each virtual departure:

[2] To prove this lemma formally, apply the rate conservation law of Miyazawa [7, Theorem 2.1 and Corollary 2.1] to the stationary process $Q_i^p R_i$.

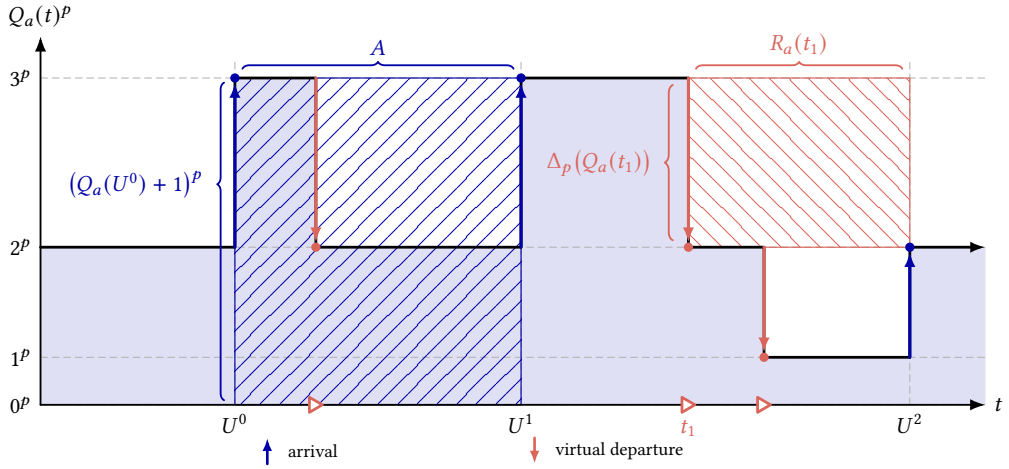


Figure 3.2. The process $Q_a(t)^p$ over two renewal cycles of the arrival process, from U^0 to U^2. Upward blue arrows mark arrivals, while downward coral arrows mark virtual departures. The blue hatched rectangle has area $A(Q_a(U^0)+1)^p$. The coral hatched cutout has area $R_a(t_1)\Delta_p\big(Q_a(t_1)\big)$, where its width is the residual time until the next arrival and its height is the decrease in Q_a^p caused by the virtual departure at t_1.

1. At the initial arrival time U^0, a base rectangle extending across the entire renewal interval, with area $A(Q_a(U^0)+1)^p$.
2. For each virtual departure time $t \in (U^0, U^1)$, a rectangular cutout extending from t until U^1, with area $R_a(t)\Delta_p\big(Q_a(t)\big)$.

The base rectangle contributes expected area

$$\mathbf{E}_a[A(Q_a+1)^p] = \mathbf{E}[A]\mathbf{E}_a[(Q_a+1)^p].$$

A cutout has expected area

$$\mathbf{E}_{a\downarrow}\big[R_a\Delta_p\big(Q_a\big)\big]$$

and occurs at rate $\lambda_{a\downarrow} = \lambda$ by Lemma 2.2. Thus, cutouts remove expected total area

$$(\lambda\mathbf{E}[A])\mathbf{E}_{a\downarrow}\big[R_a\Delta_p\big(Q_a\big)\big]$$

during one renewal interval. Combining these and applying Palm inversion (i.e., generalized renewal-reward) and Lemma 2.1 again yields the following.[3]

Lemma 3.3. *For $p \geq 1$, define $\Delta_p(q) \coloneqq q^p - (q-1)^p$ for $q \geq 1$. Then*

$$\mathbf{E}_\pi[Q^p] = \mathbf{E}_a[(Q+1)^p] - \lambda\mathbf{E}_{a\downarrow}[R_a\Delta_p\big(Q_a\big)].$$

[3]To prove this lemma formally, apply the rate conservation law of Miyazawa [7, Theorem 2.1 and Corollary 2.1] to the stationary process $Q_a^p R_a$.

We now use Lemmas 3.2 and 3.3 together to get an upper bound on $\mathbf{E}_\pi[Q^p]$ in terms of only $\mathbf{E}_{i\uparrow}[R_i^p]$, $\mathbf{E}_{a\downarrow}[R_a^p]$, p, and basic system parameters. At a token epoch of token stream i, $Q_i(0+) \le Q_i = Q$. Therefore, Lemma 3.1 gives

$$\mathbf{E}_i[Q_i(0+)^p] \le \mathbf{E}_i[Q^p] = \rho \mathbf{E}_a[(Q+1)^p].$$

Using this inequality in Lemma 3.2 gives

$$\mathbf{E}_\pi[Q^p] \le \rho \mathbf{E}_a[(Q+1)^p] + \frac{\lambda}{n}\mathbf{E}_{i\uparrow}\Big[R_i \Delta_p\big(Q_i+1\big)\Big].$$

On the other hand, Lemma 3.3 gives

$$\mathbf{E}_\pi[Q^p] = \mathbf{E}_a[(Q+1)^p] - \lambda \mathbf{E}_{a\downarrow}\Big[R_a \Delta_p\big(Q_a\big)\Big].$$

Combining these two relations and rearranging yields

$$(1-\rho)\mathbf{E}_a[(Q+1)^p] \le \frac{\lambda}{n}\mathbf{E}_{i\uparrow}\Big[R_i \Delta_p\big(Q_i+1\big)\Big] + \lambda \mathbf{E}_{a\downarrow}\Big[R_a \Delta_p\big(Q_a\big)\Big].$$

We now replace both discrete increments by simpler upper bounds. By the mean value theorem, for every $p \ge 1$ and $q \ge 1$,

$$\Delta_p(q) = q^p - (q-1)^p \le p q^{p-1}.$$

Consequently,

$$(1-\rho)\mathbf{E}_a[(Q+1)^p] \le \lambda p\left(\frac{1}{n}\mathbf{E}_{i\uparrow}\Big[R_i (Q_i+1)^{p-1}\Big] + \mathbf{E}_{a\downarrow}\Big[R_a Q_a^{p-1}\Big]\right).$$

Neither expectation on the right can generally be factored because the residual time and queue length need not be independent under the corresponding Palm laws. Hölder's inequality controls these products without requiring independence. For $p > 1$,

$$\mathbf{E}_{i\uparrow}\Big[R_i (Q_i+1)^{p-1}\Big] \le \mathbf{E}_{i\uparrow}[R_i^p]^{1/p}\mathbf{E}_{i\uparrow}[(Q_i+1)^p]^{1-1/p},$$

and

$$\mathbf{E}_{a\downarrow}\Big[R_a Q_a^{p-1}\Big] \le \mathbf{E}_{a\downarrow}[R_a^p]^{1/p}\mathbf{E}_{a\downarrow}[Q_a^p]^{1-1/p}.$$

For $p = 1$, these are both equalities, as both sides reduce to $\mathbf{E}_{i\uparrow}[R_i]$ or $\mathbf{E}_{a\downarrow}[R_a]$, respectively.

By Lemma 3.1, both queue-moment factors produced by Hölder's inequality equal $\mathbf{E}_a[(Q+1)^p]^{1-1/p}$. It follows that

$$(1-\rho)\mathbf{E}_a[(Q+1)^p] \le \lambda p\left(\frac{1}{n}\mathbf{E}_{i\uparrow}[R_i^p]^{1/p} + \mathbf{E}_{a\downarrow}[R_a^p]^{1/p}\right)\mathbf{E}_a[(Q+1)^p]^{1-1/p}.$$

Dividing by $\mathbf{E}_a[(Q+1)^p]^{1-1/p}$ gives

$$\mathbf{E}_a[(Q+1)^p]^{1/p} \le \frac{\lambda p}{1-\rho}\left(\frac{1}{n}\mathbf{E}_{i\uparrow}[R_i^p]^{1/p} + \mathbf{E}_{a\downarrow}[R_a^p]^{1/p}\right).$$

Moreover, Lemma 3.3 implies

$$\mathbf{E}_\pi[Q^p] \le \mathbf{E}_a[(Q+1)^p].$$

We have therefore proved the following.

Proposition 3.4. *For every $p \ge 1$,*

$$\mathbf{E}_\pi[Q^p] \le \mathbf{E}_a[(Q+1)^p] \le \left(\frac{\lambda p}{1-\rho}\right)^p \left(\frac{1}{n}\mathbf{E}_{i\uparrow}[R_i^p]^{1/p} + \mathbf{E}_{a\downarrow}[R_a^p]^{1/p}\right)^p.$$

4 Bounding $\mathbf{E}_{i\uparrow}[R_i^p]$ and $\mathbf{E}_{a\downarrow}[R_a^p]$

We would like to bound $\mathbf{E}_{i\uparrow}[R_i^p]$, so we consider the quantity $\lambda_{i\uparrow}\mathbf{E}_{i\uparrow}[R_i^p]$. This quantity has a simple reward-rate interpretation. Assign to each virtual arrival the reward R_i^p. Since virtual arrivals occur at rate $\lambda_{i\uparrow}$, the quantity of interest is the long-run rate at which this reward is generated.

Another way to compute the long-run reward rate is to compute the expected total reward generated during a single renewal cycle of token stream i and multiply it by μ, the renewal rate of the token stream. Under $\mathbf{P}_i[\cdot]$, time zero is a token of token stream i, and S is the time until its next token. Between these two token epochs, the virtual queue $Q_i(a)$ increases by one at each virtual arrival and has no downward jumps. It follows that[4]

$$\lambda_{i\uparrow}\mathbf{E}_{i\uparrow}[R_i^p] = \mu\mathbf{E}_i\left[\int_{(0,S)} (S-a)^p \, \mathrm{d}Q_i(a)\right]. \tag{4.1}$$

The expectation on the right has two sources of randomness:

- The length of the renewal cycle S.
- The ages a at which virtual arrivals occur during the cycle.

Intuitively, these two sources of randomness are independent because, until the next token, the virtual-arrival times depend only on the arrival process, the other token streams, and *the past of token stream i*, whereas S is the fresh renewal increment of token stream i, sampled independently of all these.

Suppose for the moment that the virtual arrivals could be represented by a process independent of S—we will soon show that this is in fact the case. After fixing $S = s$, we could then average over only the random ages at which the virtual arrivals occur, obtaining

[4]Formally, this identity is the Neveu exchange formula applied to the jointly stationary point processes of tokens from token stream i and virtual arrivals; see Baccelli and Brémaud [1, Section 1.3.2].

a deterministic age-dependent rate of virtual arrivals. Denote this rate $\nu(a)$. We would then have

$$\lambda_{i\uparrow}\mathbf{E}_{i\uparrow}[R_i^p] = \mu\mathbf{E}\left[\int_{(0,S)} (S-a)^p \nu(a)\,\mathrm{d}a\right],$$

where the expectation is now over *only the interval length* S. Note that letting $p = 0$ gets us

$$\lambda_{i\uparrow} = \mu\mathbf{E}\left[\int_{(0,S)} \nu(a)\,\mathrm{d}a\right],$$

and so

$$\mathbf{E}_{i\uparrow}[R_i^p] = \frac{\mathbf{E}\left[\int_{(0,S)} (S-a)^p \nu(a)\,\mathrm{d}a\right]}{\mathbf{E}\left[\int_{(0,S)} \nu(a)\,\mathrm{d}a\right]}. \tag{4.2}$$

Remark 4.1. When $p = 1$, Equation (4.2) gives a useful way to interpret the residual time seen by a virtual arrival. A virtual arrival samples the residual service time $S - a$ at an age a (where, as per the scheduling literature, a job's age is the amount of service it has already received) whose weight is proportional to

$$\nu(a)\mathbf{P}[S > a].$$

If $\nu(a)$ were an arbitrary nonnegative function, this sampling rule could concentrate all its weight near whichever age has the largest mean residual life. The immediate bound would then be

$$\mathbf{E}_{i\uparrow}[R_i] \le \sup_{a:\mathbf{P}[S>a]>0} \mathbf{E}[S-a \mid S > a].$$

The key additional structure is that $\nu(a)$ is non-increasing, as we will prove below. Thus, the sampling rule cannot place substantial weight at a later age without placing at least as much weight at every earlier age. This restriction leads to a substantially tighter bound. ◇

We now derive the sharper bound alluded to in Remark 4.1. For the moment, we will assume that $\nu(a)$ is non-increasing; we prove this property afterward. Because $\nu(a)$ is non-increasing, for every $y > 0$ there is a cutoff $t_y \in [0, \infty]$ such that

$$\mathbb{1}(\nu(a) > y) = \mathbb{1}(a < t_y)$$

for almost every $a \geq 0$. Using this and Tonelli's theorem,

$$\begin{aligned}\mathbf{E}\left[\int_0^S (S-a)^p \nu(a)\,\mathrm{d}a\right] &= \int_0^\infty \mathbf{E}\left[\int_0^{S\wedge t_y} (S-a)^p\,\mathrm{d}a\right]\mathrm{d}y \\ &= \int_0^\infty c_p(t_y)\mathbf{E}[S\wedge t_y]\,\mathrm{d}y \\ &\leq \sup_{t>0} c_p(t)\int_0^\infty \mathbf{E}[S\wedge t_y]\,\mathrm{d}y \\ &= \sup_{t>0} c_p(t)\mathbf{E}\left[\int_0^S \nu(a)\,\mathrm{d}a\right], \end{aligned} \tag{4.3}$$

where

$$c_p(t) \coloneqq \frac{\mathbf{E}\left[\int_0^{S\wedge t} (S-a)^p\,\mathrm{d}a\right]}{\mathbf{E}[S\wedge t]}$$

for $0 < t \leq \infty$ and $c_p(0) \coloneqq \mathbf{E}[S^p]$. Thus, (4.2) and (4.3) give

$$\mathbf{E}_{i\uparrow}[R_i^p] \leq \sup_{t>0} c_p(t).$$

It remains to bound $c_p(t)$ uniformly over $t > 0$. Since $(S-a)^p \leq S^p$ for $0 \leq a \leq S$,

$$c_p(t) = \frac{\mathbf{E}\left[\int_0^{S\wedge t} (S-a)^p\,\mathrm{d}a\right]}{\mathbf{E}[S\wedge t]} \leq \frac{\mathbf{E}[S^p(S\wedge t)]}{\mathbf{E}[S\wedge t]}.$$

We claim that

$$\frac{\mathbf{E}[S^p(S\wedge t)]}{\mathbf{E}[S\wedge t]} \leq \frac{\mathbf{E}[S^{p+1}]}{\mathbf{E}[S]} = \frac{\mathbf{E}[(\mu S)^{p+1}]}{\mu^p}.$$

To see this, observe that both ratios are weighted averages of S^p, but use different weights for an interval of length S. The first assigns weight $S\wedge t$, whereas the second assigns weight S. Passing from the first weighting to the second shifts weight toward longer intervals. Since S^p also increases with S, this shift can only increase the weighted average. Putting this all together, we obtain the following bound.

Lemma 4.2. *Fix $p \geq 1$ and suppose that $\mathbf{E}[S^{p+1}] < \infty$. Define*

$$C_{p,s} \coloneqq \sup_{t>0} \frac{\mathbf{E}\left[\int_0^{S\wedge t} (S-a)^p\,\mathrm{d}a\right]}{\mathbf{E}[S\wedge t]}.$$

Then

$$\mathbf{E}_{i\uparrow}[R_i^p] \leq C_{p,s} \leq \frac{\mathbf{E}[S^{p+1}]}{\mathbf{E}[S]} = \frac{\mathbf{E}[(\mu S)^{p+1}]}{\mu^p}.$$

To prove this bound on $\mathbf{E}_{i\uparrow}[R_i^p]$, we must close both loose ends from the above argument:

1. We must show that we can replace $Q_i(t)$ with a process independent of S in (4.1). This is done in Section 4.1.
2. We must show that $\nu(a)$ exists and is non-increasing in a. These are done in Sections 4.2 and 4.3 respectively.

We then follow similar steps to bound $\mathbf{E}_{a\downarrow}[R_a^p]$ in Section 4.4.

4.1 The stopped queue and the stopped virtual queue

Our first step is to construct an alternate process that is both independent of S and results in the same virtual arrivals on $(0, S)$. We define the *stopped queue* $\widehat{Q}(t)$ and the *stopped virtual queue* $\widehat{Q}_i(t)$. The stopped queue is constructed in the same way as $Q(t)$, except that server i is stopped at time zero: all tokens of token stream i occurring strictly after time zero are removed. Define

$$\operatorname{stop}(\vec{x}_i) \coloneqq (\dots, T_i^{-1}, T_i^0),$$

so that $\operatorname{stop}(\vec{x}_i)$ is the portion of token stream i up to and including time zero. Let $\operatorname{stop}_i(\mathbf{x})$ denote the sample point obtained by replacing $\vec{x}_i$ with $\operatorname{stop}(\vec{x}_i)$ while leaving every other primitive process unchanged. Formally,

$$\widehat{Q}(t; \mathbf{x}) = Q\Big(\operatorname{adv}\big(\operatorname{stop}_i(\mathbf{x}), t\big)\Big).$$

The construction of the stopped virtual queue $\widehat{Q}_i(t)$ is similar. In this process, server i remains stopped, but we move through virtual time by translating only token stream i, while holding the arrival process and all other token streams fixed:

$$\widehat{Q}_i(t; \mathbf{x}) = Q\Big(\operatorname{adv}_i\big(\operatorname{stop}_i(\mathbf{x}), t\big)\Big).$$

We note that, for the purpose of replacing $Q_i(t)$ in (4.1), we only needed to define the stopped virtual queue; the stopped queue $\widehat{Q}(t)$ is unnecessary. However, $\widehat{Q}(t)$ will be useful for analyzing the dynamics of $\widehat{Q}_i(t)$ in Section 4.2 due to the following fact. Under $\mathbf{P}_i[\cdot]$, the arrival process and all token streams other than token stream i retain their independent stationary laws. Advancing these processes by any fixed amount therefore does not change their joint distribution. Consequently, for every fixed $t \geq 0$,

$$\widehat{Q}(t) \overset{\mathrm{d}}{=} \widehat{Q}_i(t) \qquad \text{under } \mathbf{P}_i[\cdot]. \tag{4.4}$$

We now justify replacing $Q_i(t)$ with $\widehat{Q}_i(t)$ in (4.1). Observe that, for every $a \in (0, T_i^1)$,

$$\widehat{Q}_i(a; \mathbf{x}) = Q\Big(\operatorname{adv}_i\big(\operatorname{stop}_i(\mathbf{x}), a\big)\Big) = Q\Big(\operatorname{adv}_i\big(\mathbf{x}, a\big)\Big) = Q_i(a; \mathbf{x}).$$

The middle equality holds because, when $a < T_i^1$, none of the tokens deleted by stop_i has crossed time zero after token stream i is advanced by a. Consequently, $\operatorname{adv}_i(\operatorname{stop}_i(\mathbf{x}), a)$

and $\mathrm{adv}_i(\mathbf{x}, a)$ agree on all primitive events strictly before time zero. Since $\boldsymbol{Q}$ is determined by this history, the two configurations produce the same queue length. Thus, under $\mathbf{P}_i[\cdot]$ we may write $S = T_i^1$ and rewrite (4.1) as

$$\lambda_{i\uparrow}\mathbf{E}_{i\uparrow}[R_i^p] = \mu\mathbf{E}_i\left[\int_{(0,S)} (S-a)^p \,\mathrm{d}\widehat{Q}_i(a)\right].$$

By construction, the process $\widehat{Q}_i$ is determined by $\mathrm{stop}_i(\mathbf{x})$. Under $\mathbf{P}_i[\cdot]$, the next inter-token time $S = T_i^1$ is a fresh renewal increment and is independent of $\mathrm{stop}_i(\mathbf{x})$. Hence, $\widehat{Q}_i$ and S are independent under $\mathbf{P}_i[\cdot]$. This resolves the dependence issue we previously had in (4.1), and so we can now condition on S and use Tonelli:

$$\begin{aligned}\mathbf{E}_i\left[\int_{(0,S)} (S-a)^p \,\mathrm{d}\widehat{Q}_i(a)\right] &= \mathbf{E}_i\left[\mathbf{E}_i\left[\int_{(0,S)} (S-a)^p \,\mathrm{d}\widehat{Q}_i(a) \;\middle|\; S\right]\right] \\ &= \mathbf{E}\left[\int_{(0,S)} (S-a)^p \mathbf{E}_i\left[\mathrm{d}\widehat{Q}_i(a)\right]\right],\end{aligned}$$

where the outer expectation in the last line is taken over only the interval length S, since the expectation over the process histories has already been absorbed into $\mathbf{E}_i\left[\mathrm{d}\widehat{Q}_i(a)\right]$ and the fresh renewal increment S is independent of those histories. Thus, we have successfully replaced $Q_i(t)$ with a process independent of S in (4.1). We now must show that the measure $\mathbf{E}_i[\mathrm{d}\widehat{Q}_i(a)]$ admits an age-dependent rate function $\nu(a)$.

4.2 Determining the age-dependent virtual arrival rate $\nu(a)$

Note that $\mathrm{d}\widehat{Q}_i(a)$ is the counting measure of the virtual arrivals of $\widehat{Q}_i$, so $\mathbf{E}_i\left[\mathrm{d}\widehat{Q}_i(a)\right]$ is their mean counting measure under $\mathbf{P}_i[\cdot]$. We will identify the Radon–Nikodym derivative of this measure with respect to Lebesgue measure and denote it by $\nu(a)$. This means that $\nu(a)$ is characterized by[5]

$$\mathbf{E}_i\left[\int_{(t_1,t_2)} \mathrm{d}\widehat{Q}_i(a)\right] = \int_{t_1}^{t_2} \nu(a)\,\mathrm{d}a, \qquad 0 < t_1 < t_2,$$

so we may interpret $\nu(a)$ as the age-dependent rate of virtual arrivals.

To identify this rate, we observe that

$$\mathbf{E}_i\left[\int_{(t_1,t_2)} \mathrm{d}\widehat{Q}_i(a)\right] = \mathbf{E}_i[\widehat{Q}_i(t_2)] - \mathbf{E}_i[\widehat{Q}_i(t_1)] = \mathbf{E}_i[\widehat{Q}(t_2)] - \mathbf{E}_i[\widehat{Q}(t_1)],$$

where the second equality follows from (4.4). The dynamics of the stopped queue $\widehat{Q}(t)$ on (t_1, t_2) are as follows:

[5]We restrict to $t_1 > 0$ because the Palm token at virtual time zero may produce an initial downward jump. We therefore define ν on $(0, \infty)$; the value assigned to $\nu(0)$ does not affect any of the integrals below.

- The stopped queue increases by one at each arrival.
- At a token of token stream $j \neq i$, the stopped queue decreases by one if it is positive and remains unchanged if it is zero. No tokens from token stream i occur after time zero because server i has been stopped.

Hence, $\mathbf{E}_i[\widehat{Q}(t_2) - \widehat{Q}(t_1)]$ equals[6]

$$\mathbf{E}_i[\#\text{ of arrivals on } (t_1, t_2)] - \sum_{j \neq i} \mathbf{E}_i[\text{server } j \text{ tokens on } (t_1, t_2) \text{ that see } \widehat{Q} > 0].$$

Letting $\#(\cdot)$ denote the cardinality of a set, we can rewrite the second term as

$$\sum_{j \neq i} \mathbf{E}_i\left[\#\left(a \in (t_1, t_2) : \begin{array}{c}\text{a token of server } j \text{ occurs } a \text{ time units after} \\ \text{server } i \text{ is stopped and sees } \widehat{Q}(a) > 0\end{array}\right)\right].$$

The rate at which this event occurs for a fixed server $j \neq i$ is, intuitively

$$\mu \times \text{“}\mathbf{P}_i[\text{token arriving } a \text{ time units after server } i \text{ is stopped sees } \widehat{Q}(a) > 0]\text{”}.$$

To formalize this, for $a \geq 0$ let $\mathbf{P}_{i,j}^{(0,a)}[\cdot]$ denote the joint Palm law under which token stream i has its Palm law with a token at time zero, token stream j independently has its Palm law shifted so that its Palm token occurs at time a, and the arrival process and all remaining token streams retain their stationary laws. All primitive processes are mutually independent under this law. Thus, under $\mathbf{P}_{i,j}^{(0,a)}[\cdot]$, token streams i and j have tokens at times zero and a, respectively.

The rate of token stream j is μ, and the joint Palm probability that its token at time a sees a nonempty stopped queue is $\mathbf{P}_{i,j}^{(0,a)}[\widehat{Q}(a) > 0]$. Thus, summing over the $n-1$ servers other than i and using symmetry, we obtain

$$\begin{aligned}\mathbf{E}_i[\widehat{Q}(t_2) - \widehat{Q}(t_1)] &= \lambda(t_2 - t_1) - \mu(n-1)\int_{t_1}^{t_2} \mathbf{P}_{i,j}^{(0,a)}[\widehat{Q}(a) > 0]\,\mathrm{d}a \\ &= \int_{t_1}^{t_2} \left(\lambda - \mu(n-1)\mathbf{P}_{i,j}^{(0,a)}[\widehat{Q}(a) > 0]\right)\mathrm{d}a.\end{aligned}$$

Therefore, $\mathbf{E}_i[\mathrm{d}\widehat{Q}_i(a)]$ is absolutely continuous with intensity (Radon–Nikodym derivative)

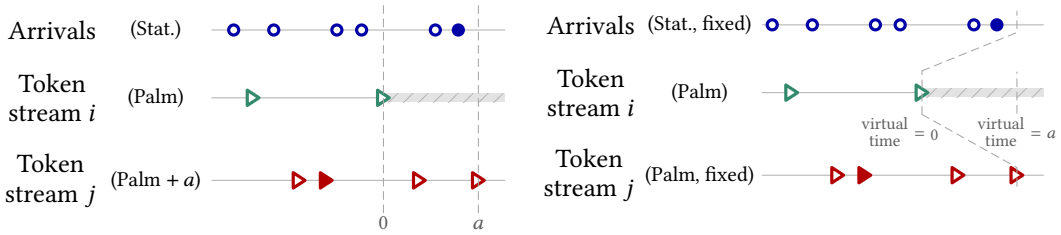
$$\nu(a) = \lambda - \mu(n-1)\mathbf{P}_{i,j}^{(0,a)}[\widehat{Q}(a) > 0] \tag{4.5}$$

for almost every $a \geq 0$.

[6] If $n = 1$, the sum below is empty, so $\mathbf{E}_i[\widehat{Q}(t_2) - \widehat{Q}(t_1)] = \lambda(t_2 - t_1)$. Thus, the mean virtual-arrival measure has density $\nu(a) = \lambda$, which is non-increasing. The remainder of the argument is therefore needed only when $n \geq 2$.

4.3 Proving the monotonicity of $v(a)$

The last step needed to prove Lemma 4.2 is to show that $v(a)$ is non-increasing in a. By (4.5), it is sufficient to prove that $\mathbf{P}_{i,j}^{(0,a)}[\widehat{Q}(a) > 0]$ is non-decreasing in a. Intuitively, this should be the case. The longer the system has operated since server i was stopped, the more tokens that server i would otherwise have supplied are missing. Since the arrivals and the other token streams are unchanged, this accumulating service deficit can only make it more likely that there are jobs waiting in the system at time a.



(a) The stopped queue $\widehat{Q}(t)$ under $\mathbf{P}_{i,j}^{(0,a)}[\cdot]$, evaluated at real time a, when the Palm token of token stream j is produced.

(b) The virtual stopped queue $\widehat{Q}_i(t)$ under $\mathbf{P}_{i,j}^{(0,0)}[\cdot]$, evaluated at virtual time a after advancing only token stream i.

Figure 4.1. Equivalent real-time and virtual-time constructions of the stopped system. At their respective evaluation times, the two panels induce the same joint law of the primitive processes.

To prove this, start by considering Figure 4.1. Each panel shows the same relative arrangement of arrivals and tokens, but reaches that arrangement in a different way:

1. Figure 4.1(a) starts under $\mathbf{P}_{i,j}^{(0,a)}[\cdot]$ and advances the stopped queue by a units of real time, from time zero to the instant at which the Palm token of token stream j is produced. This corresponds to $\mathbf{P}_{i,j}^{(0,a)}[\widehat{Q}(a) > 0]$.
2. Figure 4.1(b) starts under $\mathbf{P}_{i,j}^{(0,0)}[\cdot]$ and advances the *virtual* stopped queue for token stream i by a units of virtual time. That is, only token stream i advances while the other primitive processes remain fixed. This corresponds to $\mathbf{P}_{i,j}^{(0,0)}[\widehat{Q}_i(a) > 0]$.

To see why the two arrangements have the same joint law, compare the primitive processes at their respective evaluation times, real time a in the first construction and virtual time a in the second:

- The arrival process has its stationary law in both constructions.
- Token stream i was stopped at a Palm token a units before the evaluation time in both constructions.
- Token stream j has a Palm token at the evaluation time in both constructions.
- Every remaining token stream has its stationary law in both constructions.

Since the primitive processes are mutually independent, their joint laws at the two evaluation times agree. Therefore,

$$\mathbf{P}_{i,j}^{(0,a)}[\widehat{Q}(a) > 0] = \mathbf{P}_{i,j}^{(0,0)}[\widehat{Q}_i(a) > 0]. \tag{4.6}$$

Under the law $\mathbf{P}_{i,j}^{(0,0)}[\cdot]$, $\widehat{Q}_i(a)$ is non-decreasing in a for $a > 0$. Indeed, after the Palm token at time zero, the stopped token stream contains no further tokens, so $\widehat{Q}_i(a)$ can increase at virtual arrivals but cannot decrease. Thus, $\mathbf{P}_{i,j}^{(0,0)}[\widehat{Q}_i(a) > 0]$ is non-decreasing in a. By (4.5) and (4.6), it follows that $\nu(a)$ is non-increasing for $a > 0$, completing the proof of Lemma 4.2.

4.4 Bounding $\mathbf{E}_{a\downarrow}[R_a^p]$

We also need a bound on $\mathbf{E}_{a\downarrow}[R_a^p]$. The argument is the same as the one we just completed for $\mathbf{E}_{i\uparrow}[R_i^p]$, so we provide only a proof sketch.

Lemma 4.3. *Fix $p \geq 1$ and suppose that $\mathbf{E}[A^{p+1}] < \infty$. Define*

$$C_{p,a} \coloneqq \sup_{t>0} \frac{\mathbf{E}\left[\int_0^{A\wedge t} (A-u)^p \,\mathrm{d}u\right]}{\mathbf{E}[A \wedge t]}.$$

Then

$$\mathbf{E}_{a\downarrow}[R_a^p] \leq C_{p,a} \leq \frac{\mathbf{E}[A^{p+1}]}{\mathbf{E}[A]} = \frac{\mathbf{E}[(\lambda A)^{p+1}]}{\lambda^p}.$$

Proof sketch. Construct the stopped-arrival queue by deleting every arrival strictly after time zero while leaving all token streams unchanged. Define the corresponding virtual stopped-arrival queue by advancing only this truncated arrival process in virtual time, while holding every token stream fixed.

Under $\mathbf{P}_a[\cdot]$, the next interarrival time $A = U^1$ is a fresh renewal increment, independent of the stopped configuration. We may therefore average over the stopped configuration while holding A fixed, obtaining a deterministic rate of virtual departures at each elapsed time u after the arrival at zero. Denote this rate by $\eta(u)$. Repeating the Palm inversion (i.e., generalized renewal-reward) calculation that gets us (4.1) with reward R_a^p gives

$$\mathbf{E}_{a\downarrow}[R_a^p] = \frac{\mathbf{E}\left[\int_0^A (A-u)^p \eta(u) \,\mathrm{d}u\right]}{\mathbf{E}\left[\int_0^A \eta(u) \,\mathrm{d}u\right]}.$$

It remains to note that $\eta(u)$ is non-increasing. Indeed, the same Palm-law conversion used in (4.6) expresses the virtual-departure rate at age u using a fixed joint Palm law. Under that law, after the Palm arrival at virtual time zero, the stopped arrival process contains no further arrivals. The virtual stopped-arrival queue can therefore decrease at virtual departures but cannot increase, so its probability of being positive is non-increasing in u. The virtual-departure rate $\eta(u)$ is consequently non-increasing as well.

Applying the calculation that results in (4.3), with A replacing S and η replacing ν, gives

$$\mathbf{E}\left[\int_0^A (A-u)^p \eta(u) \,\mathrm{d}u\right] \leq C_{p,a}\mathbf{E}\left[\int_0^A \eta(u) \,\mathrm{d}u\right].$$

Substituting this inequality into the preceding ratio yields

$$\mathbf{E}_{a\downarrow}[R_a^p] \le C_{p,a}.$$

Finally,

$$C_{p,a} \le \frac{\mathbf{E}[A^p(A \wedge t)]}{\mathbf{E}[A \wedge t]} \le \frac{\mathbf{E}[A^{p+1}]}{\mathbf{E}[A]},$$

uniformly over $t > 0$, by the same weighted-average argument used for S. □

5 Main results

We now combine the results from the last three sections to prove our main results.

Theorem 1.3. *Fix $p \ge 1$ and suppose that*

$$\mathbf{E}[(\mu S)^{p+1}] + \mathbf{E}[(\lambda A)^{p+1}] < \infty.$$

Then

$$\mathbf{E}[L^p] \le \left[\frac{p}{1-\rho}\left(\rho \mathbf{E}[(\mu S)^{p+1}]^{1/p} + \mathbf{E}[(\lambda A)^{p+1}]^{1/p}\right)\right]^p.$$

Proof. The bound for $\mathbf{E}[Q^p]$ follows from Proposition 3.4 and Lemmas 4.2 and 4.3. Gamarnik and Goldberg [2, Proposition 1] proves that Q stochastically dominates L, which implies $\mathbf{E}[L^p] \le \mathbf{E}[Q^p]$. □

By slightly refining the proof, we can get a stronger bound for the mean, i.e., in the $p = 1$ case.

Theorem 1.1. *Suppose that $\rho < 1$ and*

$$\mathbf{E}[S^2] + \mathbf{E}[A^2] < \infty.$$

Then

$$\mathbf{E}[L] \le \frac{\rho\mu C_{\text{service}} + \frac{1}{2}\mathbf{E}[(\lambda A)^2] - \rho}{1-\rho} \le \frac{\rho \mathbf{E}[(\mu S)^2] + \frac{1}{2}\mathbf{E}[(\lambda A)^2] - \rho}{1-\rho}.$$

Furthermore, if the arrival process is Poisson,

$$\mathbf{E}[L] \le \frac{\rho}{1-\rho}\mu C_{\text{service}} \le \frac{\rho}{1-\rho}\mathbf{E}[(\mu S)^2].$$

Proof sketch. The first refinement is that we retain the exact effect of the token at time zero, i.e., we do not use $Q_i(0+) \le Q_i = Q$. When $p = 1$, Lemma 3.2 gives

$$\mathbf{E}_\pi[Q] = \mathbf{E}_i[Q_i(0+)] + \frac{\lambda}{n}\mathbf{E}_{i\uparrow}[R_i].$$

At a token epoch, $Q_i(0+) = \max((Q-1), 0)$, and hence

$$\begin{aligned}\mathbf{E}_i[Q_i(0+)] &= \mathbf{E}_i[Q] - \mathbf{P}_i[Q > 0] \\ &= \rho \mathbf{E}_a[Q+1] - \rho,\end{aligned}$$

where we used Lemma 3.1 and $\mathbf{P}_i[Q > 0] = \rho$. By Lemma 3.3,

$$\mathbf{E}_a[Q+1] = \mathbf{E}_\pi[Q] + \lambda \mathbf{E}_{a\downarrow}[R_a].$$

Substituting these identities and using $\lambda/n = \rho\mu$ gives

$$(1-\rho)\mathbf{E}_\pi[Q] = \rho\mu \mathbf{E}_{i\uparrow}[R_i] + \rho\lambda \mathbf{E}_{a\downarrow}[R_a] - \rho. \tag{5.1}$$

The second refinement comes from a separate rate-conservation argument for the real queue $Q(t)$, rather than for either virtual queue. Apply the rate-conservation law [7, Theorem 2.1 and Corollary 2.1] to the stationary process

$$t \mapsto R_a(t) Q(t),$$

where $R_a(t)$ is the time from real time t until the next arrival (note that previously we had only defined $R_a(t)$ for virtual time). This relates its jumps at arrivals to its jumps at useful tokens and gives a sharper representation of the arrival-residual term. Between events, $R_a(t)$ decreases at unit rate while $Q(t)$ remains constant, giving continuous drift $-\mathbf{E}_\pi[Q]$. At an arrival, the process jumps upward by $A(Q+1)$, whose contribution per unit time is

$$\lambda \mathbf{E}_a[A(Q+1)] = \mathbf{E}_a[Q+1],$$

because A is a fresh interarrival time with $\lambda \mathbf{E}[A] = 1$. At a useful token, the queue decreases by one, so the process jumps downward by R_a. The aggregate contribution of these jumps is

$$-n\mu \mathbf{E}_i[R_a \mathbb{1}(Q > 0)].$$

Rate conservation therefore gives

$$\mathbf{E}_a[Q+1] - \mathbf{E}_\pi[Q] = n\mu \mathbf{E}_i[R_a \mathbb{1}(Q > 0)].$$

Comparing this identity with Lemma 3.3 and using $\rho n \mu = \lambda$ yields

$$\rho\lambda \mathbf{E}_{a\downarrow}[R_a] = \lambda \mathbf{E}_i[R_a \mathbb{1}(Q > 0)].$$

Under $\mathbf{P}_i[\cdot]$, the arrival process retains its stationary renewal law. Therefore,

$$\lambda \mathbf{E}_i[R_a] = \frac{1}{2}\mathbf{E}[(\lambda A)^2],$$

and hence

$$\rho\lambda \mathbf{E}_{a\downarrow}[R_a] = \frac{1}{2}\mathbf{E}[(\lambda A)^2] - \mathbf{E}_i[\lambda R_a \mathbb{1}(Q = 0)]. \tag{5.2}$$

We can obtain the sharper Poisson bound at this point. If the arrival process is Poisson, then R_a is exponential with rate λ and is independent of the state of the queue under $\mathbf{P}_i[\cdot]$. Since $\mathbf{P}_i[Q = 0] = 1 - \rho$,

$$\mathbf{E}_i[\lambda R_a \mathbb{1}(Q = 0)] = 1 - \rho.$$

Moreover,

$$\frac{1}{2}\mathbf{E}[(\lambda A)^2] = 1.$$

Thus, the last two terms in (5.1) cancel:

$$\rho\lambda \mathbf{E}_{a\downarrow}[R_a] - \rho = 0.$$

Hence,

$$(1 - \rho)\mathbf{E}_\pi[Q] = \rho\mu \mathbf{E}_{i\uparrow}[R_i],$$

which gives the stated $M/G/n$ bound after applying Lemma 4.2.

For general renewal arrivals, substitute (5.2) into (5.1) and apply Lemma 4.2 to obtain

$$(1 - \rho)\mathbf{E}_\pi[Q] \le \rho\mu C_{\text{service}} + \frac{1}{2}\mathbf{E}[(\lambda A)^2] - \rho - \mathbf{E}_i[\lambda R_a \mathbb{1}(Q = 0)].$$

Dropping the final nonnegative term gives the simpler $GI/GI/n$ bound. Finally, Lemma 4.2 with $p = 1$ gives

$$\mu C_{\text{service}} \le \mathbf{E}[(\mu S)^2]. \qquad \square$$

The above result is even stronger if S is increasing mean residual life (IMRL).

Corollary 1.2. *If S is IMRL, the bound in Theorem 1.1 becomes*

$$\mathbf{E}[L] \le \frac{\rho\mathbf{E}[(\mu S)^2] + \mathbf{E}[(\lambda A)^2] - 2\rho}{2(1 - \rho)}.$$

If, in addition, the arrival process is Poisson, then

$$\mathbf{E}[L] \le \frac{\rho}{2(1 - \rho)}\mathbf{E}[(\mu S)^2].$$

To prove this, we first need to prove the following simple lemma, which evaluates C_{service} when S is IMRL.

Lemma 5.1. *Suppose that $\mathbf{E}[S^2] < \infty$ and that S is IMRL. Then*

$$\mu C_{\text{service}} = \frac{1}{2}\mathbf{E}[(\mu S)^2].$$

Proof. For every $a \geq 0$ such that $\mathbf{P}[S > a] > 0$, define

$$m_1(a) \coloneqq \mathbf{E}[S - a \mid S > a].$$

By Tonelli's theorem,

$$\frac{\mathbf{E}\left[\int_0^{S \wedge t} (S - a)\, \mathrm{d}a\right]}{\mathbf{E}[S \wedge t]} = \frac{\int_0^t m_1(a)\mathbf{P}[S > a]\, \mathrm{d}a}{\int_0^t \mathbf{P}[S > a]\, \mathrm{d}a}.$$

Since S is IMRL, $m_1(a)$ is non-decreasing, and hence these weighted prefix averages are non-decreasing in t. Therefore,

$$C_{\text{service}} = \lim_{t \to \infty} \frac{\mathbf{E}\left[\int_0^{S \wedge t} (S - a)\, \mathrm{d}a\right]}{\mathbf{E}[S \wedge t]} = \frac{\mathbf{E}[S^2]}{2\mathbf{E}[S]}.$$

Multiplying by $\mu = 1/\mathbf{E}[S]$ gives

$$\mu C_{\text{service}} = \frac{\mu^2 \mathbf{E}[S^2]}{2} = \frac{1}{2}\mathbf{E}[(\mu S)^2]. \qquad \square$$

Corollary 1.2 then follows immediately from Lemma 5.1 and Theorem 1.1.

6 Extensions for planned preprint

There are a number of related results that we intend to include in the soon-to-be-completed preprint. These are results that we mostly have proofs for but left out of this note to keep it simple and to the point. These include:

- A strengthening of the p-moment bound that improves it in certain regimes.
- A generalization of the result to a setting where the arrival process is a superposition of m mutually independent renewal processes.
- A generalization of the result to a setting where the servers are heterogeneous.
- Results when $p < 1$.
- A discussion of how tight these bounds are, and in particular how tight they are for the *modified* $GI/GI/n$ construction. This would point towards needing a completely new proof approach to close the gap between these bounds and the single-server Kingman bound (assuming doing so is possible).

References

[1] François Baccelli and Pierre Brémaud. *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences*. Number 26 in Stochastic Modelling and Applied Probability. Springer, Berlin, Germany, 2 edition, 2003. ISBN 978-3-540-66088-0. doi: 10.1007/978-3-662-11657-9.

[2] David Gamarnik and David A Goldberg. Steady-state gi/gi/n queue in the halfin-whitt regime. *The Annals of Applied Probability*, pages 2382–2419, 2013.

[3] Yige Hong. A new $1/(1-\rho)$-scaling bound for multiserver queues via a leave-one-out technique, October 2025.

[4] Aleksandr Khintchine. Mathematische Theorie der stationären Reihe [Mathematical theory of a stationary queue]. *Matematicheskii Sbornik*, 39(4):73–84, 1932. URL `http://mi.mathnet.ru/msb6834`.

[5] J. F. C. Kingman. Some inequalities for the queue GI/G/1. *Biometrika*, 49(3/4):315, December 1962. ISSN 00063444. doi: 10.2307/2333966.

[6] Yuan Li and David A. Goldberg. Simple and explicit bounds for multiserver queues with $1/(1-\rho)$ scaling. *Mathematics of Operations Research*, 50(2):813–837, May 2025. ISSN 0364-765X, 1526-5471. doi: 10.1287/moor.2022.0131.

[7] Masakiyo Miyazawa. Rate conservation laws: A survey. *Queueing Systems*, 15(1):1–58, March 1994. ISSN 1572-9443. doi: 10.1007/BF01189231.

[8] Felix Pollaczek. Über eine Aufgabe der Wahrscheinlichkeitstheorie. I [On a problem in probability theory. I]. *Mathematische Zeitschrift*, 32(1):64–100, December 1930. ISSN 1432-1823. doi: 10.1007/BF01194620.

[9] Felix Pollaczek. Über eine Aufgabe der Wahrscheinlichkeitstheorie. II [On a problem in probability theory. II]. *Mathematische Zeitschrift*, 32(1):729–750, December 1930. ISSN 1432-1823. doi: 10.1007/BF01194663.