# Slay Your Delays

## Scheduling and Dispatching in Queues

Ziv Scully

Cornell ORIE 7590, Spring 2025

*Last updated 2025-07-27 18:32*

# Contents

# Chapter 1
# Analyzing work via Palm calculus

The *work* in a queueing system is the total remaining time needed to serve all jobs currently in the system. We model the work in a system as a stationary process $W = (t \mapsto W(t))$. In the case of a single-server queue, $W$ has relatively simple behavior [§ 1.2].

- When the server is busy, namely when $W(t) > 0$, the work $W(t)$ has slope $-1$.
- When the server is idle, namely when $W(t) = 0$, the work $W(t)$ has slope 0.
- At certain times, specifically *arrival times* of jobs, the work jumps up according to the size of the arriving job, i.e. a job that arrives at time $t$ has size $W(t+) - W(t-)$.

See Figure 1.1 for an illustration of $W$ in a single-server queue.

The purpose of this chapter is to develop tools for answering the following question.

*Question 1.1.* What factors affect the stationary amount of work in a queueing system?

The main tool we introduce for studying this question is *Miyazawa's Rate Conservation Law (RCL)* from *Palm calculus* [§ 1.3]. The RCL gives us relationships between different "forces", like service and arrivals, that act on the work in a queueing system. One obstacle is that some forces, like service, act continuously; while others, like arrivals, act via discontinuous jumps. Palm calculus gives us a language for working with jumps.

Using the RCL and Palm calculus, we'll be able to answer Question 1.1 for several types of queueing systems. For a few simple models, namely the single-server *M/G/1 queue* [§ 1.2] and variants thereof [Exrs. 1.13, 1.14], we obtain exact formulas. For more complex models, such as the single-server *G/G/1 queue* [§ 1.4.3] and the multiserver *M/G/k queue* [Exr. 1.15], we obtain incomplete formulas that still provide useful bounds, approximations, or other insights.

## 1.1 Foundations: stationarity and point processes

This section defines stationary processes and introduces the notation and terminology conventions we use when working with them.

### 1.1.1 General conventions

*Notation 1.2.*

(a) The notation $t \mapsto \mathrm{expr}[t]$ denotes the function (or stochastic process) that maps $t$ to $\mathrm{expr}[t]$, which is some expression involving $t$. This is useful for describing one-off functions or processes without giving them a name. For instance, we can write $t \mapsto t^2$ instead of "$f$, where $f(t) = t^2$". We emphasize that the variable used is arbitrary, e.g. $(t \mapsto t^2) = (u \mapsto u^2)$, but we'll do our best to reduce confusion by choosing variables that don't conflict with others.

(b) When it can be done without introducing ambiguity, we combine functions with arithmetic operations, e.g. $(X + Y)(t) = X(t) + Y(t)$, $(X^2)(t) = X(t)^2$, etc.

*Notation 1.3.* We (shamelessly!) commit several probability sins.

(a) The words "almost surely" will be omitted almost everywhere.

(b) Independence is understood to be *mutual* independence unless otherwise specified.

(c) "Process" is understood to mean "stochastic process", i.e. random functions, typically with domain $\mathbb{R}$ or $\mathbb{Z}$. Ordinary deterministic functions are, of course, a special case. The only exception is the term *point process* [Def. 1.10], which is a random set rather than a random function.

(d) Rather than explicitly defining measures, we define their expectation operators, which we denote by $\mathbf{E}[\cdot]$ with various subscripts and superscripts [Def. 1.20], and we never work directly with the elements of any underlying sample space. We write the corresponding probability measure as $\mathbf{P}[\cdot]$ with the same subscripts and superscripts.

(e) When introducing a random variable, we use the term *fresh* or *freshly* to indicate it is independent of all previously defined random variables.

(f) Throughout, we work with one "main" probability space with expectation $\mathbf{E}[\cdot]$. Abusing notation somewhat, we also use $\mathbf{E}[\cdot]$ for "one-off" expectations. However, all other probabilities and expectations use a subscript and/or superscript.

(g) A notation $X$ may stand for either a distribution or a fresh random variable with that distribution, i.e. "$X \sim X$". Such a random variable is either defined on its own "one-off" probability space, or, if it's defined on the main probability space, is independent of everything else in the expression.

(h) For the most part, distributions of random variables are understood to be under the main probability measure $\mathbf{P}[\cdot]$. When discussing distributions under a different measure $\mathbf{P}'[\cdot]$, we say so explicitly.

(i) Measurability serves essentially two purposes in probability theory: preventing pathological examples (e.g. the Vitali set is not Borel-measurable) and discussing information (e.g. $Y$ is $X$-measurable if you can figure out the value of $Y$ if you know the value of $X$). We implicitly assume the former type of measurability throughout, reserving explicit discussion of measurability for the latter.

**Assumption 1.4.** Unless otherwise specified, all stochastic processes $t \mapsto X(t)$ on domain $\mathbb{R}$ are piecewise differentiable with finite derivatives, possibly with jump discontinuities [Def. 1.11] at which both the left limit $X(t-)$ and right limit $X(t+)$ exist. We assume the mean number of jump discontinuities occurring in any compact interval is finite.

**Definition 1.5.**

(a) The *uniform distribution* $\mathrm{Unif}(R)$ refers the uniform distribution with respect to either the counting measure (if $R$ is finite) or the Lebesgue measure (if $R \subset \mathbb{R}^n$ is uncountable and compact). For example, $\mathbf{P}[\mathrm{Unif}[0,1) < t] = t$ for $t \in [0,1]$.

(b) The *exponential distribution* $\mathrm{Exp}(\lambda)$ with parameter $\lambda > 0$ is defined by, for $t \geq 0$,

$$\mathbf{P}[\mathrm{Exp}(\lambda) > t] = e^{-\lambda t}.$$

(c) The *geometric distributions* $\mathrm{Geo}_0(p)$ and $\mathrm{Geo}_1(p)$ with parameter $p \in [0, 1]$ are defined by, for *integer* $i \geq 0$,

$$\mathbf{P}[\mathrm{Geo}_0(p) \geq i] = \mathbf{P}[\mathrm{Geo}_1(p) > i] = (1 - q)^i.$$

The subscript 0 or 1 indicates whether the minimum possible value is 0 or 1.

*Notation 1.6.* We sometimes drop parentheses around function arguments that already have their own delimiters. For instance, we write $\mathrm{Unif}[0, 1) := \mathrm{Unif}([0, 1))$ in Definition 1.5(a).

### 1.1.2 Stationary processes

**Definition 1.7.**

(a) The *shifting* of a process $X$ by $t \in \mathbb{R}$ is the process $X_{\mathrm{shift}}(t)$ defined by

$$(X_{\mathrm{shift}}(t))(u) := X(t + u).$$

Intuitively, $X_{\mathrm{shift}}(t)$ is what $X$ would be if we renamed "time $t$" to "time 0".

(b) A process $X$ is *stationary* if the distribution of $X_{\mathrm{shift}}(t)$ (as a random function) is the same for all $t$. This is equivalent to saying that for all $u_1, \ldots, u_n \in \mathbb{R}$, the joint distribution of

$$X(t + u_1), \ldots, X(t + u_n)$$

is the same for all $t$.

(c) More generally, processes $X_1, \ldots, X_n$ are *jointly stationary* if the distribution of $((X_1)_{\mathrm{shift}}(t), \ldots, (X_n)_{\mathrm{shift}}(t))$ (as a tuple of random functions) is the same for all $t$. This is equivalent to saying that for all $u_1, \ldots, u_n \in \mathbb{R}$, the joint distribution of

$$X_1(t + u_1), X_1(t + u_2), \ldots, X_n(t + u_{m-1}), X_n(t + u_m)$$

is the same for all $t$. This means that if $X$ is stationary, then any $m$ of its shiftings $X_{\mathrm{shift}}(u_1), \ldots, X_{\mathrm{shift}}(u_m)$ are jointly stationary.

*Notation 1.8.*

(a) We typically call the input of a process "time", or sometimes "index" when the domain is $\mathbb{Z}$. We usually use $t$ and nearby letters as the input variable when the domain is $\mathbb{R}$, and we typically use $i$ and nearby letters when the domain is $\mathbb{Z}$.

(b) When discussing a stationary process $X$ at a generic time $t$ whose value is not important, we often pick $t = 0$ for concreteness, and to reduce clutter, we often write

$$X := X(0),$$

though we sometimes write out $X(0)$ for clarity. In particular, we ensure that it's clear from context whether $X$ refers to the process $t \mapsto X(t)$ or the random variable $X(0)$.

(c) Similarly to (b), to discuss left and right limits at time $t = 0$ with less clutter, we write

$$X_{(-)} := X(0-),$$
$$X_{(+)} := X(0+).$$

(d) In informal discussion, e.g. in Remark 1.19, we discuss time as measured in "seconds" as opposed to the less committal but less poetic "time units". This choice is, of course, arbitrary and without loss of generality.

*Remark 1.9.* The most important intuition to have about a stationary process $X$ is the following. Imagine sampling a specific path $X$, then freshly sampling a random time $T$ from any distribution. Because $X$ is stationary and independent of $T$, the distribution of $X(T)$ is the same as the distribution of $X(t)$ for any fixed $t$. The takeaway is that *whenever we look at a stationary process's value, we can imagine we're viewing it at a random time.*

### 1.1.3 Point processes and jumps

We model the work in a queueing system as a stationary process $W$ [§ 1.2]. Our main goal in this chapter is to understand its stationary distribution. To do so, it helps to have notation and terminology for discussing the times at which $W$ jumps, namely arrival times.

**Definition 1.10.** A *point process* $A \subset \mathbb{R}$ is a random countable set of points.

(a) $N_A(R) := \#(A \cap R)$ is the number of points from $A$ in $R$.

(b) $\lambda_A := \mathbf{E}[N_A(0, 1]]$ is the *rate* of the point process.

(c) $A_i(t)$ is the $i$th element of $A$, where elements are sorted, indexed by $i \in \mathbb{Z}$, and

$$\ldots < A_{-1}(t) < A_0(t) \le t < A_1(t) < \ldots.$$

For example, $A_1(t) := \min(A \cap (t, \infty))$ is the first point of $A$ (strictly) after time $t$, and $A_0(t) = t$ if and only if $t \in A$.

(d) The *shifting* of $A$ is the set-valued process

$$A_{\text{shift}}(t) := \{a - t : a \in A\} = \{A_i(t) - t : i \in \mathbb{Z}\}.$$

Intuitively, $A_{\text{shift}}(t)$ is what $A$ would be if we renamed "time $t$" to "time 0".

We call a point process $A$ *stationary* if its shifting $A_{\text{shift}}$ is stationary, and similarly for $A$ being *jointly stationary* with any set of other processes.

**Definition 1.11.** Let $X$ be a process (satisfying Assumption 1.4). We define the following processes to discuss the continuous and discontinuous motion of $X$.

(a) The *(right) derivative* of $X$ is the process

$$\mathrm{D}X(t) := \lim_{\delta \to 0+} \frac{X(t + \delta) - X(t+)}{\delta}.$$

(b) The *jump magnitudes*, or simply *jump*, of $X$ is the process

$$\Delta X(t) := X(t+) - X(t-).$$

(c) The *jump times* of $X$ is the point process

$$\mathrm{J}X := \{t \in \mathbb{R} : \Delta X(t) \neq 0\}.$$

One can check that if $X$ is stationary, then $X$, $\mathrm{D}X$, $\Delta X$, and $\mathrm{J}X$ are jointly stationary.

*Notation 1.12.* We introduce analogues of D, Δ, and J that operate on expressions. Specifically, if $t$ is a variable and $\mathrm{expr}[t]$ is an expression involving $t$, then

$$\begin{aligned} \mathrm{D}_t \,\mathrm{expr}[t] &:= \big(\mathrm{D}(u \mapsto \mathrm{expr}[u])\big)(t), \\ \Delta_t \,\mathrm{expr}[t] &:= \big(\Delta(u \mapsto \mathrm{expr}[u])\big)(t), \\ \mathrm{J}_t \,\mathrm{expr}[t] &:= \mathrm{J}(u \mapsto \mathrm{expr}[u]). \end{aligned}$$

For instance, $\mathrm{D}_t$ is the same as the usual $\frac{\mathrm{d}}{\mathrm{d}t}$ derivative notation, except we specify that we refer to the right derivative.

Given the work process $W$ of a queue, its jump times $\mathrm{J}W$ represent times when work is added to the queue, usually due to arriving jobs.

## 1.2 The M/G/1 queue

For the types of queues we study, there are three main questions we need to answer to define a queueing model.

- When do jobs arrive to the system?
- What does each job look like? In particular, what is each job's *size*, i.e. how much time does each job take to serve?
- How are jobs served once they are in the system?

We'll start by considering single-server queues, specifically focusing on analyzing the amount of *work* in the queue, which is the total amount of time it would take to finish all jobs currently in the queue. We focus on work for now because simplifies the story for the last question: we don't need to worry about exactly which job is in service, because serving any job will decrease the system's total work at the same rate. As long as the server stays busy whenever there is work to be done, the work process obeys the dynamics described in Definition 1.13 below.

**Definition 1.13.** A nonnegative process $W \geq 0$ is a *standard work process* if for all $t$:

(a) $\mathrm{D}W(t) = -1 + \mathbb{1}(W(t) = 0)$, i.e. it decreases at rate 1 whenever it is nonzero.

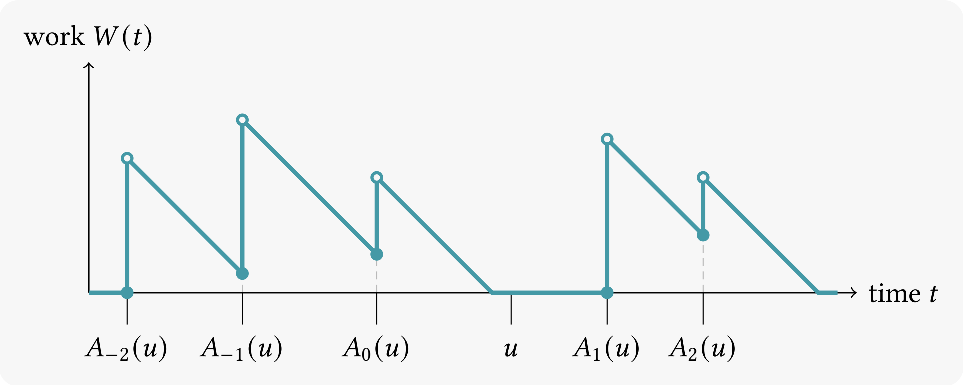


**Figure 1.1.** Work $W(t)$ in a single-server queueing system as a function of time $t$. Work is measured in time it takes to complete, so the slope is $-1$ whenever the system is nonempty, i.e. whenever $W(t) > 0$. Arrivals correspond to upwards jumps, where the arriving job's size is the height of the jump. Discontinuities are drawn as vertical lines for clarity. The arrival times are $A_i(u)$, where the index $i \in \mathbb{Z}$ indicates position relative to an arbitrarily chosen fixed time $u$.

(b) $\lambda_{JW} > 0$ and $\Delta W(t) \geq 0$, i.e. it has positive jumps.[1]

(c) $W(t) = W(t-)$, i.e. it is *left-continuous*. We use this convention because we often want to discuss the state of the system as observed by a job "just before" it arrives, but our use for "just after" is less common. With that said, for jump times $t \in JW$, we often write $W(t-)$ to emphasize the jump.

More generally, $W$ is a *work process* under the same conditions, but with (b) relaxed to $DW(t) \in [-1, 0]$.

See Figure 1.1 for an illustration of a standard work process $W$ with arrival times $A = JW$. For now, we focus on queueing models with standard work processes. As you'll encounter in Exercise 1.15, non-standard work processes are much harder to analyze.

To fully define a standard work process $W$, we need to define how work arrives, which amounts to defining $\Delta W$. Here we introduce *M/G arrivals* [Def. 1.15], which strikes a balance between theoretical tractability and modeling flexibility.

**Definition 1.14.** A stationary point process $A$ is a *(homogeneous) Poisson process* if any of the following equivalent conditions hold.

(a) In $\mathbf{E}[\cdot]$, the following sequence is i.i.d. with distribution $\text{Exp}(\lambda_A)$:

$$\ldots, (A_{-1} - A_{-2}), (A_0 - A_{-1}), (0 - A_0), (A_1 - 0), (A_2 - A_1), (A_3 - A_2), \ldots \sim \text{Exp}(\lambda_A).$$

(b) In $\mathbf{E}_A[\cdot]$, meaning $A_0 = 0$, the following sequence is i.i.d. with distribution $\text{Exp}(\lambda_A)$:

$$\ldots, (A_{-1} - A_{-2}), (A_0 - A_{-1}), (A_1 - A_0), (A_2 - A_1), \ldots \sim \text{Exp}(\lambda_A).$$

[1] There is no special reason to require $\lambda_{JW} > 0$, but the $\lambda_{JW} = 0$ edge case is trivial, so we allow ourselves the liberty of dividing by $\lambda_{JW}$.

(c) In either $\mathbf{E}[\cdot]$ or $\mathbf{E}_A[\cdot]$, the process $A$ has *independent increments*, meaning that for any set of disjoint sets $R_1, \dots, R_n$, the random variables $N_A(R_1), \dots, N_A(R_n)$ are mutually independent.

**Definition 1.15.** A *$(\lambda, S)$-M/G arrival process* (or simply *M/G arrivals*) with *arrival rate* $\lambda > 0$ and *size distribution* $S > 0$ is a process of the form given below.

(a) Let the *arrival times* $A$ be a fresh Poisson process of rate $\lambda_A = \lambda$.

(b) Let the *job sizes* be a fresh sequence $\dots, S_{-1}, S_0, S_1, \dots \sim S$.

(c) Let the *load* be $\rho := \lambda\,\mathbf{E}[S]$.

Then the process

$$S(t) := \begin{cases} S_i & \text{if } t = A_i \\ 0 & \text{otherwise} \end{cases}$$

is an M/G arrival process. That is, $S(t) \sim S$ freshly if $t \in A$, and $S(t) = 0$ otherwise.

**Definition 1.16.**

(a) We say a work process $W$ *has M/G arrivals* if its jump $\Delta W$ is an M/G arrival process. A *(standard) $(\lambda, S)$-M/G/1 work process* is a standard work process that has M/G arrivals.

We have not yet defined the “M/G/1 queue” itself. For now, because of our focus on work [Q. 1.1], we consider “M/G/1” as synonymous with “M/G/1 work process”, but we will later study aspects beyond work.

*Notation 1.17.*

(a) When discussing M/G arrivals or an M/G/1, we use the notation in Definition 1.15 by default.

(b) As we have done already, we omit the “$(\lambda, S)$-” in front of “M/G” unless we need to disambiguate or otherwise emphasize the parameters $\lambda$ and $S$. We do similarly for other definitions with prefixes throughout [Def. 1.31].

(c) We overload the letter $S$ in Definition 1.15 because for the most part, we rarely need the $S$-based notation aside from “$S$” by itself, standing for the size distribution or the size of one arriving job. In particular, if an arrival happens at time $t = 0$, then the size $S_0 = S(0)$ of the arrival is a fresh sample from $S$, which we usually denote by simply $S$ [Ntn. 1.3(g)].

(d) While (c) covers most cases, we do occasionally discuss the $i$th size $S_i$. To be able to do, we need $S_i$ to arise as the value at time 0 of a stationary process, so we let

$$S_i(t) := S(A_i(t)),$$

which, consistent with Notation 1.8(b), satisfies $S_i = S_i(0)$.

(e) When discussing multiple systems with M/G arrivals, such as when comparing variants of the M/G/1 to the standard M/G/1 [Exrs. 1.13–1.15], by default, all the M/G arrival processes being discussed have the same arrival rate and size distribution.

(f) When defining processes involving i.i.d. samples in the future, we give brief descriptions like "$S(t) \sim S$ freshly if $t \in A$" without explicitly defining the sequence of samples from $S$. See Definition 1.31 for an example.

**Assumption 1.18.** Unless otherwise stated, we consider only M/G arrivals with $\rho < 1$.

*Remark 1.19.* Our requirement that the load satisfy $\rho < 1$ is to ensure *stability* of the system, namely the existence of a stationary M/G/1 work process. We will discuss stability more formally later on, but here is the rough idea.

One can view $\rho := \lambda\,\mathbf{E}[S]$ as the average seconds of work added to the queue per second: roughly speaking, $\lambda$ jobs per second bring an average of $\mathbf{E}[S]$ seconds of work each. If $\rho > 1$, then on average, more work arrives per second than the server can complete in one second. This means work tends to increase over time on average, which means the work process can't possibly be stationary. Less obviously, instability also occurs when $\rho = 1$, in analogy with null recurrent Markov chains.

If $\rho < 1$, as we usually assume, then it also has another interpretation: it is the fraction of time an M/G/1 is busy [Exr. 1.5].

## 1.3 Palm calculus and three of its key formulas

To analyze a stationary work process $W$, we need to be able to write statements related to how $W$ changes over time. Working with the derivative $\mathrm{D}W$ is relatively easy, but the jump $\Delta W$ is more difficult, because we can't condition on a jump occurring at any fixed time $t$. Specifically, because $\mathrm{J}W$ is countable [Asm. 1.4], stationarity implies $\mathbf{P}[t \in \mathrm{J}W] = 0$.

*Palm calculus* is a set of techniques that allows us to "condition on a jump happening at time $t$" without dividing by zero. It takes the form of a key definition [§ 1.3.1] and several key formulas about it, three of which we focus on here [§§ 1.3.2–1.3.4].

### 1.3.1 Defining Palm expectation

**Definition 1.20.** Let $X \geq 0$, a nonnegative process, and $A$, a point process, be jointly stationary. The *Palm expectation* of $X$ with respect to stationary point process $A$ at time $t$ is

$$\mathbf{E}_A^t[X(t)] := \frac{1}{\lambda_A u}\,\mathbf{E}\left[\sum_{a \in A \cap (0,u]} X(a)\right] = \frac{1}{\lambda_A u}\,\mathbf{E}\left[\sum_{i=1}^{N(0,u]} X(A_i)\right],$$

where $u > 0$ is arbitrary, as all values of $u$ yield the same result by stationarity [Exr. 1.1]. Similarly, the time $t$ is arbitrary, so we often omit it [Ntn. 1.8(b)], writing

$$\mathbf{E}_A[X] = \mathbf{E}_A^t[X(t)].$$

For general $X$, we define $\mathbf{E}_A[X] = \mathbf{E}_A[X^+] - \mathbf{E}_A[X^-]$, which means $\mathbf{E}_A[X]$ is undefined if $\mathbf{E}_A[X^+] = \mathbf{E}_A[X^-] = \infty$.

As discussed in Notation 1.3(d), we denote the probability measure corresponding to $\mathbf{E}_A^t[\cdot]$, called the *Palm probability* with respect to $A$ at time $t$, by $\mathbf{P}_A^t[\cdot] = \mathbf{E}_A^t[\mathbb{1}(\cdot)]$.

The key intuition one should have about Palm expectation is

$$\mathbf{E}_A^t[X(t)] = \text{“}\,\mathbf{E}[X(t) \mid t \in A]\text{”},$$

where the right-hand side isn't rigorously defined because $\mathbf{P}[t \in A] = 0$. But the above intuition concisely communicates a lot of things about $\mathbf{E}_A^t[\cdot]$. The most important of these is that $\mathbf{E}_A^t[\cdot]$ is a valid expectation operator, representing integration with respect to a valid probability measure $\mathbf{P}_A^t[\cdot]$, so all the usual properties of expectations (e.g. Markov's inequality) hold for $\mathbf{E}_A^t[\cdot]$.

*Remark 1.21.* If $X$ and $A$ are jointly stationary, then $\mathbf{E}_A^t[X(s)]$ makes sense even if $s \neq t$. The key is to view $X(s)$ as the value of the stationary process $r \mapsto X(r+s-t)$ at time $t$, so

$$\mathbf{E}_A^t[X(s)] = \frac{1}{\lambda_A u}\,\mathbf{E}\left[\sum_{a \in A \cap (0,u]} X(a+s-t)\right] = \frac{1}{\lambda_A u}\,\mathbf{E}\left[\sum_{i=1}^{N(0,u]} X(A_i+s-t)\right].$$

### 1.3.2 RCL: Miyazawa's Rate Conservation Law

**Theorem 1.22: Miyazawa's Rate Conservation Law (RCL).** *Let $X$ be a stationary process.*

*(a) We have, possibly with $+\infty$ on both sides,*[2]

$$\mathbf{E}[(\mathrm{D}X)^+] + \lambda_{\mathrm{J}X}\,\mathbf{E}_{\mathrm{J}X}[(\Delta X)^+] = \mathbf{E}[(\mathrm{D}X)^-] + \lambda_{\mathrm{J}X}\,\mathbf{E}_{\mathrm{J}X}[(\Delta X)^-].$$

*(b) If $\mathbf{E}[|\mathrm{D}X|] < \infty$ and $\mathbf{E}_{\mathrm{J}X}[|\Delta X|] < \infty$, then*

$$\mathbf{E}[\mathrm{D}X] + \lambda_{\mathrm{J}X}\,\mathbf{E}_{\mathrm{J}X}[\Delta X] = 0.$$

*Both conclusions still hold if $\mathrm{J}X$ is replaced by a point process $A \supseteq \mathrm{J}X$ that is jointly stationary with $X$.*

Miyazawa's Rate Conservation Law is important enough to deserve a brief but distinct name. For the benefit of those reading in hyperlinkless print, we call it “RCL 1.22”. We do similarly for other key formulas throughout. See Baccelli and Brémaud [1, Section 1.3.3] and Miyazawa [7] for more about RCL 1.22, its proof, and its generalizations.

[2] We write $x^+ = \max\{x, 0\}$ and $x^- = (-x)^+$ for the positive and negative parts of $x$, respectively.

*Remark 1.23.*

(a) Written out for a generic time $t$, RCL 1.22(a) says

$$\mathbf{E}\big[(\mathrm{D}_t X(t))^+\big] + \lambda_{\mathrm{J}X}\,\mathbf{E}^t_{\mathrm{J}X}\big[(\Delta_t X(t))^+\big] = \mathbf{E}\big[(\mathrm{D}_t X(t))^-\big] + \lambda_{\mathrm{J}X}\,\mathbf{E}^t_{\mathrm{J}X}\big[(\Delta_t X(t))^-\big],$$

and, also writing out $\Delta_t X(t)$ [Def. 1.11(b)], RCL 1.22(b) says,

$$\mathbf{E}[\mathrm{D}_t X(t)] + \lambda_{\mathrm{J}X}\,\mathbf{E}^t_{\mathrm{J}X}[X(t+) - X(t-)] = 0.$$

(b) The fact that one may replace $\mathrm{J}X$ with any stationary point process $A \supset \mathrm{J}X$ follows from Exercise 1.2.

(c) RCL 1.22(b) can hold even if $\mathbf{E}[|X|] = \infty$, and this is useful in many applications of it. However, it never holds when $\mathbf{E}[|\mathrm{D}X|] = \infty$ or $\mathbf{E}_{\mathrm{J}X}[|\Delta X|] = \infty$, because an implicit conclusion is that $\mathbf{E}[\mathrm{D}X]$ and $\mathbf{E}_{\mathrm{J}X}[\Delta X]$ are well defined and finite.

(d) Baccelli and Brémaud [1, Section 1.3.3] state and prove RCL 1.22(b) under the stronger precondition that $|X|$ is bounded, but they note that it holds under the weaker precondition we give in RCL 1.22(b) [1, Remark 1.3.4]. In fact, they point out that if $\mathbf{E}[|X|] < \infty$, then *either* $\mathbf{E}[|\mathrm{D}X|] < \infty$ or $\mathbf{E}_{\mathrm{J}X}[|\Delta X|] < \infty$ implies the other.

(e) Miyazawa [7] gives a survey of more general RCLs. Both parts of RCL 1.22 arise as special cases of his results [7, Theorem 2.1, Remark 2.2]. RCL 1.22(a) is especially useful, as it does not require checking finiteness of expectations.

### Applying Miyazawa's RCL to the M/G/1

We can try to apply RCL 1.22 to analyze $\mathbf{E}[W]$ for an M/G/1 work process $W$. A general rule of thumb is that to understand $\mathbf{E}[f(W)]$, one should apply RCL 1.22 with $X = g(W)$, where $g$ is roughly the integral of $f$. So let's apply RCL 1.22(b) to $X = W^2$.

- For the derivative, using Definition 1.15, we compute[3]

$$\begin{aligned}\mathrm{D}X(t) &= \mathrm{D}_t W(t)^2 \\ &= 2W(t)\cdot \mathrm{D}W(t) \\ &= 2W(t)\big(-1 + \mathbb{1}(W(t) = 0)\big) \\ &= -2W(t).\end{aligned}$$

- For the jump, if $t \in A = \mathrm{J}X$ is an arrival time, then using Definition 1.15, we compute

$$\begin{aligned}\Delta X(t) &= \Delta_t W(t)^2 \\ &= W(t+)^2 - W(t-)^2 \\ &= (W(t-) + S)^2 - W(t-)^2 \\ &= 2SW(t-) + S^2,\end{aligned}$$

where $S$, the size of the arriving job [Ntn. 1.17(c)], is independent of $W(t-)$.

[3] As usual, $\mathrm{D}_t W(t)^2$ means $\mathrm{D}_t(W(t)^2)$, not $(\mathrm{D}_t W(t))^2$. We use the same convention for $\Delta_t$.

Assuming for now that $\mathbf{E}[S^2] < \infty$ to ensure the precondition of RCL 1.22(b), we get

$$\begin{aligned} 0 &= -\mathbf{E}[2W] + \lambda\,\mathbf{E}_A[2SW_{(-)} + S^2] \\ &= -2\,\mathbf{E}[W] + 2\rho\,\mathbf{E}_A[W_{(-)}] + \lambda\,\mathbf{E}[S^2]. \end{aligned} \tag{1.1}$$

To clarify, the second line follows by

- the independence of $S$ from $W_{(-)}$ [Def. 1.15(b)],
- the fact that $\rho = \lambda\,\mathbf{E}[S]$ [Def. 1.15(c)], and
- our convention for expectations with "one-off" samples from $S$ [Ntn. 1.3(f)].

Unfortunately, we are now at an impasse: we have a potential formula for $\mathbf{E}[W]$, but it involves $\mathbf{E}_A[W_{(-)}]$. We can think of $\mathbf{E}_A[W_{(-)}]$ as the mean amount of work observed by an arriving job (excluding its own size). This seems at least as difficult to characterize as the stationary mean $\mathbf{E}[W]$. Fortunately, for the special case of M/G arrivals, we are in luck! We will see in Section 1.3.4 that, somewhat miraculously,

$$\mathbf{E}_A[W_{(-)}] = \mathbf{E}[W], \tag{1.2}$$

from which we obtain

$$\mathbf{E}[W] = \frac{\frac{\lambda}{2}\,\mathbf{E}[S^2]}{1-\rho} = \frac{\frac{1}{2}(1+c_S^2)\rho}{1-\rho}\,\mathbf{E}[S], \tag{1.3}$$

where $c_S^2 = \mathbf{Var}[S]/\mathbf{E}[S]^2$ is the squared coefficient of variation of $S$. We further discuss this formula and provide an intuitive interpretation of it in Section 1.4.2.

Of course, we could have written $W$ instead of $W_{(-)}$ throughout, thanks to our left-continuity convention [Def. 1.13(c)]. We will often do this in the future, but we were explicit about jumps for this first derivation.

*Remark 1.24.* Strictly speaking, (1.3) does not immediately follow from (1.1) and (1.2), because $\mathbf{E}[W] = \infty$ is also a possible solution. One can rule this out using a truncation argument: instead of using RCL 1.22 on $W^2$, we use it on $(\min\{W, m\})^2$, obtain an upper bound on $\mathbf{E}[W\,\mathbb{1}(W \le m)]$, then take the $m \to \infty$ limit to show $\mathbf{E}[W]$ is finite.

We usually omit such finiteness-verifying truncation arguments in order to stay focused on the main queueing theory ideas. But see the proof of Lemma 2.22 for an example of such a truncation argument worked out in full. A rule of thumb is that if you can use RCL 1.22 on some process $Y$ to show a concrete upper bound on an expectation $\mathbf{E}[X]$ assuming only $\mathbf{E}[X] < \infty$ a priori, then one can use RCL 1.22 on a truncated process $\min\{Y, n\}$ to bound $\mathbf{E}[X\,\mathbb{1}(X \le m)]$ (where $m$ and $n$ are related but not necessarily the same), then take the $m \to \infty$ limit to show $\mathbf{E}[X] < \infty$.

### 1.3.3 PIF: Palm Inversion Formula

**Theorem 1.25: Palm Inversion Formula (PIF).** *Let $X \ge 0$, a nonnegative stationary process, and $A$, a point process, be jointly stationary. Then*

$$\mathbf{E}[X] = \lambda_A\,\mathbf{E}_A\left[\int_0^{A_1} X(u)\,\mathrm{d}u\right].$$

*Proof.* We apply RCL 1.22 to

$$Y(t) := \int_t^{A_1(t)} X(u)\,\mathrm{d}u. \tag{1.4}$$

We need to understand the derivative and jump of $Y$.

- The derivative is
$$\mathrm{D}Y(t) = -X(t).$$
- $Y$ jumps only when $A_1$ jumps, which happens at points in $A$, so $\mathrm{J}Y \subseteq A$.
- For $t \in A$, by right-continuity of $A_1$ [Def. 1.10(c)], we have
$$A_1(t-) = t,$$
$$A_1(t+) = A_1(t),$$
and thus
$$\Delta Y(t) = Y(t+) - Y(t-) = \int_t^{A_1(t)} X(u)\,\mathrm{d}u - 0.$$

Combining this with RCL 1.22 tells us

$$\mathbf{E}[X(t)] = -\mathbf{E}[\mathrm{D}Y(t)] = \lambda_A\,\mathbf{E}_A^t[\Delta Y(t)] = \lambda_A\,\mathbf{E}_A^t\left[\int_t^{A_1(t)} X(u)\,\mathrm{d}u\right]. \qquad \square$$

*Remark 1.26.* It is worth checking carefully that $Y$ from (1.4) is indeed jointly stationary with $X$ and $A$. One way to see this is to rewrite $Y(t)$ using shiftings [Defs. 1.7(a), 1.10(d)]:

$$Y(t) = \int_0^{A_1(t)-t} X(t+u)\,\mathrm{d}u = \int_0^{(A_{\text{shift}}(t))_1} (X_{\text{shift}}(t))(u)\,\mathrm{d}u.$$

Having defined $Y(t)$ in terms of $X_{\text{shift}}(t)$ and $A_{\text{shift}}(t)$, and knowing already that $X$ and $A$ are jointly stationary, we can conclude that $Y$ is jointly stationary with $X$ and $A$, too. In more detail:

- By joint stationarity of $X$ and $A$, the joint distribution of $X_{\text{shift}}(t), A_{\text{shift}}(t)$ is the same for all $t$.
- Because shiftings of a stationary process are jointly stationary, for all $u_1, \ldots, u_n$, the joint distribution of
$$X_{\text{shift}}(t+u_1), \ldots, X_{\text{shift}}(t+u_m), A_{\text{shift}}(t+u_1), \ldots, A_{\text{shift}}(t+u_n)$$
is the same for all $t$.
- Because $Y(t)$ and $X(t)$ are functions of $A_{\text{shift}}(t)$ and $X_{\text{shift}}(t)$, for all $u_1, \ldots, u_n$, the joint distribution of
$$Y(t+u_1), \ldots, Y(y+u_n), X(t+u_1), \ldots, X(t+u_m), A_{\text{shift}}(t+u_1), \ldots, A_{\text{shift}}(t+u_n)$$
is the same for all $t$. And this is the same as joint stationarity of $Y$, $X$, and $A_{\text{shift}}$, as desired.

### 1.3.4 PASTA: Poisson Arrivals See Time Averages

**Theorem 1.27: Poisson Arrivals See Time Averages (PASTA).** *Let $X \geq 0$, a nonnegative process (satisfying Assumption 1.4), and $A$, a Poisson process, be jointly stationary. Suppose $\{X(u) : u < t\}$ is independent of $\{a \in A : a \geq t\}$ for all times $t$.*[4] *Then*

$$\mathbf{E}[X] = \mathbf{E}_A[X_{(-)}].$$

*In particular, if $X$ is left-continuous, then*

$$\mathbf{E}[X] = \mathbf{E}_A[X].$$

See Wolff [8] or Baccelli and Brémaud [1, Section 3.3.1] for a proof of PASTA 1.27, or have a go yourself [Exrs. 1.9–1.11]. Because an M/G/1 work process evolves in a way that is independent of when future arrivals occur [Defs. 1.13, 1.16], PASTA 1.27 implies (1.2), the last missing ingredient behind the M/G/1 $\mathbf{E}[W]$ formula (1.3).

*Remark 1.28.* Written out for a generic time $t$, PIF 1.25 says

$$\mathbf{E}[X(t)] = \lambda_A\, \mathbf{E}_A^t\left[\int_t^{A_1(t)} X(u)\,\mathrm{d}u\right],$$

and PASTA 1.27 says

$$\mathbf{E}[X(t)] = \mathbf{E}_A^t[X(t-)].$$

*Remark 1.29.* We can relax the $X \geq 0$ constraint of PIF 1.25 and PASTA 1.27 to $X^-$ being bounded. We can remove the constraint entirely under some conditions on $X$ [8]. But we'll make do with nonnegative processes.

## 1.4 Work ≈ intensity × variability

In this section, we finally return to Question 1.1 in earnest. We'll see that as a rule of thumb, there are two main properties of the arrival process that determine the stationary work distribution, both of which appear in our M/G/1 $\mathbf{E}[W]$ formula (1.3).

- *Intensity* refers broadly to (a factor related to) how *quickly* work is added to the system. In (1.3), this manifests as the factor $\frac{\lambda}{1-\rho}$.[5]
- *Variability* refers broadly to (a factor related to) how *irregularly* work is added to the system. In (1.3), this manifests as the factor $\mathbf{E}[S^2] = \mathbf{Var}[S] + \mathbf{E}[S]^2$.

[4] By stationarity, if this holds for one time $t$, e.g. $t = 0$, then it holds for all times $t$.

[5] We introduce the term "intensity" to distinguish the formally defined load $\rho$ [Def. 1.15] from the vague idea of intensity referred to in the rule of thumb. While intensity often manifests in work formulas as a factor in the dominant term (hence the "×" in the rule of thumb), this intensity factor is seldom simply $\rho$. One should not confuse this vague notion of intensity with the formally defined *intensity of a point process* [1, Section 1.8], which we won't cover here.

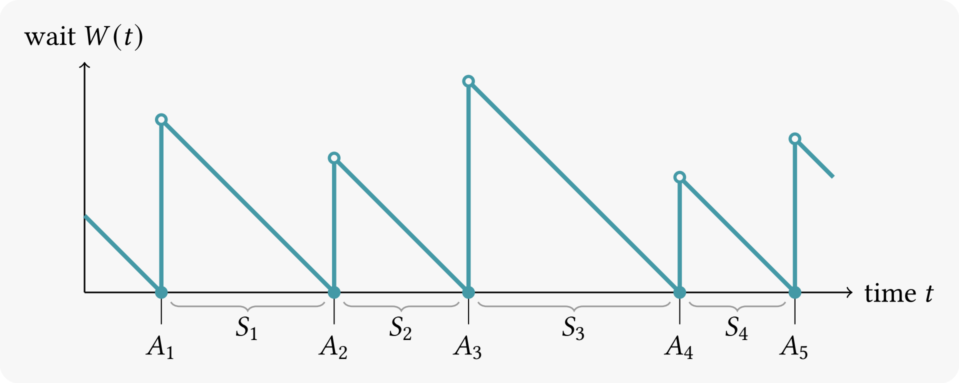


**Figure 1.2.** An $S$-renewal process $A$ [Def. 1.30] and its bus process $W$ [Def. 1.31]. This models a bus stop where the gaps between the bus arrival times $A$ are fresh samples from $S$. The wait $W(t)$ is how long you would have to wait for the next bus if you arrived at time $t$. We write $S_i \coloneqq \Delta W(A_i)$ for the gap between the $i$th and $(i+1)$th arrivals.

### 1.4.1 Effect of job size variance: waiting for the bus

**Definition 1.30.** Let $S > 0$ be a distribution. An *$S$-renewal process* with gap distribution $S$ is a point process $A$ such that *conditional on* $A_0$, the following sequence is i.i.d. with distribution $S$:

$$\ldots, (A_{-1} - A_{-2}), (A_0 - A_{-1}), (A_1 - A_0), (A_2 - A_1), \ldots \sim S.$$

Provided $\mathbf{E}[S] < \infty$, one can show that a stationary $S$-renewal process exists, so we restrict attention to this case.

**Definition 1.31.**

(a) Let $A$ be a point process. The *bus process of $A$* (a.k.a. *forward recurrence process of $A$*) is a standard work process $W$ defined by

$$W(t) = A_1(t) - t.$$

The intuition is that if $A$ is the times buses arrive to a bus stop, then $W(t)$, which we call the *wait* at $t$, is the time from $t$ until the next bus.

(b) If $S$ is a distribution on $(0, \infty)$, then an *$S$-bus process* is the bus process of an $S$-renewal process.

See Figure 1.2 for an illustration of an $S$-renewal process and its bus process.

**Theorem 1.32.** *Let $W$ be a stationary $S$-bus process.*

*(a) The mean wait is*

$$\mathbf{E}[W] = \frac{\mathbf{E}[S^2]}{2\,\mathbf{E}[S]}.$$

*(b) The transform of wait is, for $\theta \in \mathbb{R}$,*[6]

$$\mathbf{E}[e^{\theta W}] = \frac{\mathbf{E}[e^{\theta S}] - 1}{\theta\,\mathbf{E}[S]}.$$

*(c) The tail probability of wait is, for $x \geq 0$,*

$$\mathbf{P}[W > x] = \frac{\mathbf{E}[(S-x)^+]}{\mathbf{E}[S]}.$$

*Proof.* One can derive (a) and (b) from (c), but it is instructive to prove each of them directly. For simplicity, we assume that the right-hand sides are finite in these direct proofs of (a) and (b). One could extend them to cover the infinite case by following the same strategy as Exercise 1.12(b).

We prove all of the parts using RCL 1.22 on functions of $W$. Writing $A := JW$ for the $S$-renewal process [Def. 1.31], we need to understand $\mathrm{D}W$, $\lambda_A$, and the joint distribution of $W_{(-)}$ and $W_{(+)}$ under $\mathbf{P}_A[\cdot]$.

- By Definitions 1.13(a) and 1.31, $\mathrm{D}W = -1$.
- By Definition 1.31, $W_{(-)} = 0$ and $W_{(+)} = \Delta W \sim S$ freshly under $\mathbf{P}_A[\cdot]$.
- Applying RCL 1.22 to $W$ and using the above facts yields

  $$0 = -1 + \lambda_A\,\mathbf{E}_A[W_{(+)} - W_{(-)}] = -1 + \lambda_A\,\mathbf{E}[S],$$

  so $\lambda_A = 1/\mathbf{E}[S]$.

With the above in hand, we just need to apply RCL 1.22 to the right processes.

(a) Applying RCL 1.22 to $W^2$ yields

$$0 = \mathbf{E}[2W \cdot \mathrm{D}W] + \lambda_A\,\mathbf{E}_A[W_{(+)}^2 - W_{(-)}^2] = -2\,\mathbf{E}[W] + \frac{\mathbf{E}[S^2]}{\mathbf{E}[S]}.$$

(b) Applying RCL 1.22 to $e^{\theta W}$ yields

$$0 = \mathbf{E}[\theta e^{\theta W} \cdot \mathrm{D}W] + \lambda_A\,\mathbf{E}_A[e^{\theta W_{(+)}} - e^{\theta W_{(-)}}] = -\theta\,\mathbf{E}[e^{\theta W}] + \frac{\mathbf{E}[e^{\theta S}] - 1}{\mathbf{E}[S]}.$$

(c) Applying RCL 1.22 to $(W-x)^+$ yields

$$\begin{aligned} 0 &= \mathbf{E}[\mathbb{1}(W > x) \cdot \mathrm{D}W] + \lambda_A\,\mathbf{E}_A[(W_{(+)} - x)^+ - (W_{(-)} - x)^+] \\ &= -\mathbf{P}[W > x] + \frac{\mathbf{E}[(S-x)^+]}{\mathbf{E}[S]}. \end{aligned}$$ □

[6] We use the term "transform" as a catch-all for moment generating function, Laplace transform, probability generating function, etc. For $\theta \in \mathbb{R}$, the transform $\mathbf{E}[e^{\theta X}]$ of $X$ is always well defined, though it may be $\infty$.

**Definition 1.33.** Let $X \geq 0$ be a distribution with $\mathbf{E}[X] \in (0, \infty)$. The *(stationary) excess* of $X$, denoted $X_{\mathrm{e}}$, is the distribution on $(0, \infty)$ defined by

$$\mathbf{P}[X_{\mathrm{e}} > x] = \frac{\mathbf{E}[(X-x)^+]}{\mathbf{E}[X]} = \frac{1}{\mathbf{E}[X]} \int_x^\infty \mathbf{P}[X > y] \, \mathrm{d}y.$$

One can check that, in line with Theorem 1.32,

$$\mathbf{E}[X_{\mathrm{e}}^p] = \frac{\mathbf{E}[S^{p+1}]}{(p+1)\,\mathbf{E}[S]}, \qquad \mathbf{E}[e^{\theta X_{\mathrm{e}}}] = \frac{\mathbf{E}[e^{\theta X}] - 1}{\theta\,\mathbf{E}[S]}.$$

*Remark 1.34.* In brief, Theorem 1.32 says that for a stationary $S$-bus process $W$, at all times $t$,

$$W(t) \sim S_{\mathrm{e}}.$$

But because $W$ is stationary, we can interpret $W(t)$ as $W$ viewed at a random time [Rmk. 1.9]. This gives us the following intuition for $S_{\mathrm{e}}$: if we observe an $S$-renewal process at a random time, then the *remaining time* until the the next point is distributed as $S_{\mathrm{e}}$.

We will see excess distributions pop up in many settings where we care about the remaining time for something to happen, such as the remaining work of a job in an M/G/1 queue [§ 1.4.2].

What does our study of bus processes say about Question 1.1? The clearest relationship comes from viewing a bus process as the work in a queue that *always has exactly one job*. Specifically, the “job” is “wait for the next bus to arrive”, so the size distribution is $S$. The presence of an $\mathbf{E}[S^2]$ in Theorem 1.32(a) tells us that even in a queue with only one job, there can be a large amount of work on average if the *job size variance* is high. We can make this more precise by rewriting $\mathbf{E}[W] = \mathbf{E}[S_{\mathrm{e}}]$ in terms of $\mathbf{Var}[S]$:

$$\mathbf{E}[S_{\mathrm{e}}] = \frac{\mathbf{E}[S^2]}{2\,\mathbf{E}[S]} = \frac{\mathbf{Var}[S] + \mathbf{E}[S]^2}{2\,\mathbf{E}[S]^2}\,\mathbf{E}[S] = \frac{1 + c_S^2}{2}\,\mathbf{E}[S], \tag{1.5}$$

where $c_S^2 = \mathbf{Var}[S]/\mathbf{E}[S]^2$ is the squared coefficient of variation of $S$.

*Remark 1.35.* In Definition 1.31, we asserted that a stationary $S$-renewal process exists provided $\mathbf{E}[S] < \infty$. One would hope that Theorem 1.32 and Definition 1.33 would essentially tell us how to do this. Specifically, it might seem like freshly sampling $A_1 \sim S_{\mathrm{e}}$ then freshly sampling the gaps $A_{i+1} - A_i \sim S$ would make $A$ a stationary $S$-renewal process.

However, there is an issue: the above sampling procedure might result in $A_0 > 0$. The above procedure still works “forwards in time”, sampling $A \cap (0, \infty)$. Similarly, by symmetry, one could sample $A \cap (-\infty, 0]$ by going in reverse, starting with $A_0 \sim -S_{\mathrm{e}}$. However, $A_0$ and $A_1$ are in general *not* independent, so we cannot simply combine the above procedures. For example, when $S = s$ is deterministic, $A_1 = s - A_0 \sim \mathrm{Unif}(0, s]$. We need to find the *joint distribution* of $A_0$ and $A_1$ that makes the process stationary when the rest of the gaps are freshly sampled from $S$. You will figure out what this distribution must be in Exercise 1.17.

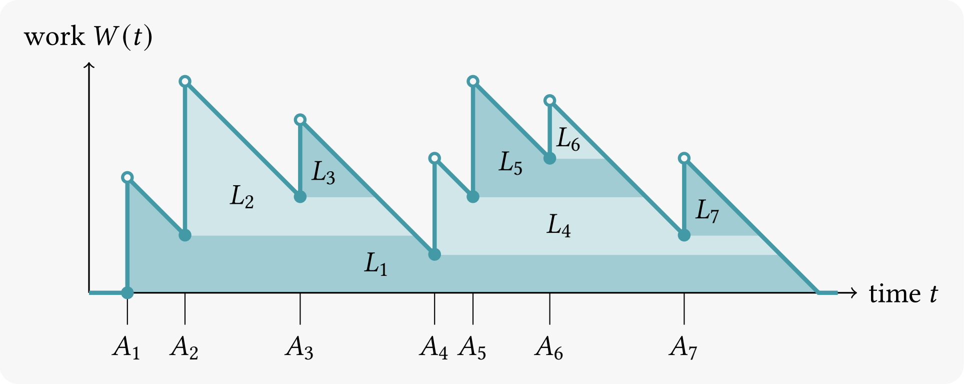


**Figure 1.3.** The "work layers" of an M/G/1 work process $W$. The layer labeled $L_i$ corresponds to the job that arrives at time $A_i$. The height of a job's layer at time $t$ is what that job's remaining work would be under the *Preemptive Last-Come First-Served (PLCFS)* policy, which always serves the job that arrived most recently. Put another way, the height of the top layer decreases at rate 1, other layer heights stay put, and each arrival creates a new layer on top of existing ones.

### 1.4.2 Effect of load: M/G/1 vs. waiting for the bus

We continue exploring Question 1.1 by looking at the mean work in an M/G/1. We saw already at the start of Section 1.4 that the mean work depends on both the job size variance and the load. In this section, we will gain further insight into the M/G/1 by comparing it to a bus process using the same distribution $S$. To that end, let

- $W_{\text{bus}}$ be an $S$-bus process and
- $W_{\text{M/G/1}}$ be a $(\lambda, S)$-M/G/1 work process.

To understand the relationship between $\mathbf{E}[W_{\text{bus}}]$ and $\mathbf{E}[W_{\text{M/G/1}}]$, it helps to express them using formulas that look as similar as possible. With the definition of $S_{\text{e}}$ in hand [Def. 1.33], we can rewrite the mean formulas from Theorem 1.32(a) and (1.3) as

$$\mathbf{E}[W_{\text{bus}}] = \mathbf{E}[S_{\text{e}}], \qquad \mathbf{E}[W_{\text{M/G/1}}] = \frac{\rho}{1-\rho}\mathbf{E}[S_{\text{e}}].$$

This way of writing the formulas makes it seem as if the M/G/1 is secretly a pile of $\frac{\rho}{1-\rho}$ buses stacked on top of each other. Remarkably, there is a sense in which this is true! Below, we give a mostly intuitive argument, with references to exercises throughout that fill in some of the gaps. In particular, see Exercise 1.18 for a way to formalize the end result.

To find the $S$-buses hiding inside a $(\lambda, S)$-M/G/1, it helps to draw the M/G/1 work process in a way that emphasizes individual job sizes, because those are the quantities that are sampled from $S$. Figure 1.3 illustrates this by associating with each job a "layer" of work. To understand these layers, it helps to think of the M/G/1 as using a scheduling policy called *PLCFS*, explained below. But this is not a restriction: all scheduling policies that don't leave the server unnecessarily idle result in the same M/G/1 work process $W_{\text{M/G/1}}$, so we are free to imagine that it is PLCFS.

The *Preemptive Last-Come First-Served (PLCFS)* scheduling policy, in brief, always serves the *job that arrived most recently*. The "preemptive" in the name comes from the fact that whenever a new job arrives, if a job is in service, we *preempt* that job in service so that we can immediately start serving the new job. We are assuming that preemption is instant and that all progress on the preempted job is retained.

### Warmup: the top layer, i.e. analyzing the job in service under PLCFS

Let's start by thinking about just the top layer, imagining the system uses PLCFS scheduling so that we can discuss this in terms of jobs. To understand the first layer, we ask: what is the chance there is a job in service; and if so, how much remaining work does the job in service have?

- As long as the server has work, there is at least one job in service, which happens with probability $\mathbf{P}[W > 0] = \rho$ [Exr. 1.5]
- Given that there is at least one job, the intuition from Remark 1.34 tells us that we can imagine that we have observed that job "at a random time", and that the job's remaining work is distributed as the excess $S_\mathrm{e}$ [Exr. 1.16(f)].

Combining these observations tells us

$$\begin{aligned} \mathbf{P}[\text{there is a job in service with remaining work} > x] &= \rho\,\mathbf{P}[S_\mathrm{e} > x], \\ \mathbf{E}[\text{remaining work of job in service (or 0 if no jobs)}] &= \rho\,\mathbf{E}[S_\mathrm{e}]. \end{aligned}$$

### Number of layers, i.e. number of jobs under PLCFS

More generally, we can ask: under PLCFS, what is the distribution of the number of jobs $N$; and given that a job is present, what is its remaining work distribution?

Perhaps surprisingly, the distribution of $N$ is pretty simple: it's geometric! Specifically, $N \sim \mathrm{Geo}_0(1-\rho)$ [Exr. 1.18(c)]. The easiest way to show this is to argue that for all $k \in \mathbb{N}$,

$$\mathbf{P}[N \geq k+1 \mid N \geq k] = \rho, \tag{1.6}$$

which implies $\mathbf{P}[N \geq k] = \rho^k$. We do not yet have the tools to give a short rigorous argument of this, but here is a short sketch. We already know $\mathbf{P}[W = 0] = 1-\rho$ and $\mathbf{P}[W > 0] = \rho$ [Exr. 1.5], which means $\mathbf{P}[N = 0] = 1-\rho$ and $\mathbf{P}[N \geq 1] = \rho$. Imagine removing all times $t$ when $N(t) = 0$ from the timeline, which amounts to conditioning on $\mathbf{P}[N \geq 1]$. The resulting system looks just like an ordinary M/G/1 using PLCFS, except there's always an extra job on the bottom layer. This means

$$\mathbf{P}[N \geq 2 \mid N \geq 1] = \mathbf{P}[N \geq 1] = \rho.$$

The story for (1.6) is essentially the same, except we condition on $N \geq k$ instead of $N \geq 1$, which yields an M/G/1 using PLCFS with $k$ extra jobs at the bottom instead of just 1.

### Height of each layer, i.e. remaining work of each job under PLCFS

We now turn to the remaining work of each job. Above, we used Remark 1.34 to argue that the job in service has remaining work distributed as $S_\mathrm{e}$. Perhaps surprisingly, the same is true of every other job, and independently so! That is, conditional on $N$, the remaining work amounts $R_1, \dots, R_N$ are i.i.d. with distribution $S_\mathrm{e}$ [Exr. 1.18(d)].

To sketch why this is, let's start by thinking about the bottom layer. Let $R_1(t)$ be the remaining work in the job that arrived longest ago that is present at time $t$, with $R_1(t) = 0$ if the system is empty at $t$. The key idea is that we can view $R_1$ as a *sometimes-paused* $S$-bus process. Specifically, in terms of an $S$-bus process $W_\mathrm{bus}$ and the number of jobs $N$, we can write, at least for $t \geq A_0$,

$$R_1(t) = \mathbb{1}(N(u) \geq 1)\, W_\mathrm{bus}\left(A_0 + \int_{A_0}^{t} \mathbb{1}(N(u) = 1)\, \mathrm{d}u\right).$$

That is, if we filter out the times $t$ when $N(t) = 0$, we can view $R_1$ as an $S$-bus process that advances while $N(t) = 1$, pauses when $N(t) \geq 2$, then resumes when $N(t) = 1$ again.

The question is thus: do the pauses introduce any differences between the stationary distributions of $R_1$ and $W_\mathrm{bus}$? Because the arrivals are Poisson, the pauses happen at "uniformly random" times. Additionally, while the the pause times are random, they are independent of when in the timeline they happen. (Specifically, one can show that each pause is essentially a freshly sampled *busy period* [Exr. 1.8].) This should convince us that the steady-state distribution is the same as that of $W_\mathrm{bus}$, namely $S_\mathrm{e}$, so we conclude

$$\mathbf{P}[R_1 > x \mid N \geq 1] = \mathbf{P}[W_\mathrm{bus} > x] = \mathbf{P}[S_\mathrm{e} > x].$$

The above story led us to $\mathbf{P}[R_1 > x \mid N \geq 1]$ by thinking about the bottom layer. If we think about the $k$th layer from the bottom instead, we find $\mathbf{P}[R_k \mid N \geq k] = \mathbf{P}[S_\mathrm{e} > x]$ by essentially the same reasoning. To show that all the layers are independent, we have to think about multiple layers at a time. For example, we might reason about pausing the bottom two layers while $N(t) \geq 3$, arguing that the length of the pause is independent of the events $R_1(t) > x_1$ and $R_2(t) > x_2$. But writing this down precisely is unwieldy enough that the approach outlined in Exercise 1.18, which requires only applying RCL 1.22 sufficiently cleverly, starts to look more appealing.

## 1.4.3 Effect of arrival time variance: G/G/1 vs. M/G/1

As a last example, we will see how variability in the arrival time process impacts the work process. To do this, we first need to define a queueing model where we can vary the arrival time variability. The *G/G/1 queue*, defined below, serves as such a model.

**Definition 1.36.**

(a) An *S-GD/G arrival process* of an *arrival times* point process $A$ (or simply *GD/G arrivals* of $A$) with *size distribution* $S > 0$ is a process $t \mapsto S(t)$ such that $S(t) \sim S$ freshly if $t \in A$, and $S(t) = 0$ otherwise.

(b) A *$(R, S)$-G/G arrival process* (or simply *G/G arrivals*) with *interarrival time distribution* $R > 0$ is a G/G arrival process whose arrival times are an $R$-renewal process (that is independent of the job sizes).

Just like Definition 1.15, we also define the *arrival rate* $\lambda := \lambda_A$ and *load* $\rho := \lambda\,\mathbf{E}[S]$. For G/G arrivals, we can also write the arrival rate as $\lambda = 1/\mathbf{E}[R]$ [Exr. 1.4].

**Definition 1.37.**

(a) We say a work process $W$ *has GD/G arrivals* if its jump $\Delta W$ is a GD/G arrival process, and analogously for saying $W$ *has G/G arrivals*.

(b) A *(standard) GD/G/1 work process* is a standard work process that has GD/G arrivals, and analogously for a *(standard) G/G/1 work process*.

*Notation 1.38.* The conventions of Notation 1.17 apply, plus we let $R_i(t) := A_{i+1}(t) - A_i(t)$ be the gap between the $i$th and $(i+1)$th arrival times.

*Notation 1.39.* Now that we are moving beyond just M/G arrivals, it is worth explaining the $\cdot/\cdot$ notation for arrival processes. The first slot describes the arrival times, and the second slot describes the job sizes. Each slot, can contain any of the following:

(a) D, for *deterministic*, indicates interarrival times or job sizes are deterministic.

(b) M, for *memoryless*, indicates interarrival times or job sizes are generated by fresh sampling from an exponential distribution [Def. 1.5(b)]. For arrivals, this corresponds to arrival times being a Poisson process [Def. 1.14].

(c) G, for *general (independent)*, indicates interarrival times or job sizes are generated by fresh sampling from a positive distribution. By default, we call the distribution $S$ for

(d) GD, for *general dependent*, indicates only that the arrival time point process or job size sequence is stationary.

Actually, in the literature, our "G" is denoted "GI", and our "GD" is denoted "G". We reserve the shorter "G" for the independent case because we use it far more commonly.

Queueing systems are typically notated using their arrival process, followed by a third slot with the number of servers, and sometimes followed by an additional slot with more information. For instance, the "1" in M/G/1 or G/G/1 refers to the fact that there is a single server, whereas you will study the $k$-server M/G/$k$ in Exercise 1.15.

This type of notation is called *Kendall notation* [4].

### Analyzing mean work in the G/G/1 seen by arrivals

Let $W$ be a stationary G/G/1 work process. We will actually analyze $\mathbf{E}_A[W] = \mathbf{E}_A[W_{(-)}]$, the mean work *seen by arrivals*, instead of $\mathbf{E}[W]$, but it is not too hard to express $\mathbf{E}[W]$ and $\mathbf{E}_A[W]$ in terms of each other [Exr. 1.19]. Throughout this section, we take advantage of the left-continuity of $W$, writing $W$ and $W(t)$ instead of $W_{(-)}$ and $W(t-)$ to reduce clutter.

Our plan is to use a discrete-time analogue of the RCL that you will prove in Exercise 1.6. In our setting, it says, roughly, saying that if we take the perspective of a job arriving at time $A_0$, then the work we see, namely $W(A_0)$, has the same distribution as the work the following job sees, namely $W(A_1)$. Specifically, applying Exercise 1.6 to the process $f(W)$ and point process $A$, we get that as long as $\mathbf{E}_A[|f(W)|] < \infty$,

$$\mathbf{E}_A[f(W(A_0))] = \mathbf{E}_A[f(W(A_1))]. \tag{1.7}$$

Also recall $W(A_0) = W$ [Ntn. 1.8(b)], so the left-hand-side is $\mathbf{E}_A[f(W)]$.

In order to use (1.7), we need to understand the relationship between $W(A_0)$ and $W(A_1)$. Remembering that $W$ is a standard work process [Def. 1.13, Fig. 1.1], we see that there are three things that can affect how $W$ changes between (just before) $A_0$ and (just before) $A_1$.

- At time $A_0$, a job of size $S_0 = S(A_0)$ arrives.
- During the time interval $(A_0, A_1)$, we serve up to $A_1 - A_0$ work.
- The only thing that might stop us from serving $A_1 - A_0$ work is if the work hits zero prior to time $A_1$. If this happens, then the work will still be zero at time $A_1$.

Combining these observations yields

$$W(A_1) = \big(W(A_0) + S_0 - (A_1 - A_0)\big)^+.$$

a recursion known as the *Lindley equation*. Plugging this into (1.7) and writing $R$ and $S$ for variables that are fresh samples from the corresponding distributions [Ntn. 1.3(g)] (namely $S_0$ and $(A_1 - A_0)$, respectively), we get

$$\mathbf{E}_A[f(W)] = \mathbf{E}_A\big[f\big((W + S - R)^+\big)\big]. \tag{1.8}$$

We know the drill now: apply (1.8) with just the right function $f$. The "try the integral" rule of thumb for RCL 1.22 applies just as well to (1.7) and (1.8). Choosing $f(x) = x^2$ yields, assuming $\mathbf{E}[R^2] < \infty$ and $\mathbf{E}[S^2] < \infty$,

$$\begin{aligned}
\mathbf{E}_A[W^2] &= \mathbf{E}_A\Big[\big((W + S - R)^+\big)^2\Big] \\
&= \mathbf{E}_A[(W + S - R)^2] - \mathbf{E}_A\Big[\big((R - S - W)^+\big)^2\Big] \\
&= \mathbf{E}_A[W^2] + 2\,\mathbf{E}_A[W]\,\mathbf{E}[S - R] + \mathbf{E}[(S - R)^2] - \mathbf{E}_A\Big[\big((R - S - W)^+\big)^2\Big],
\end{aligned}$$

where the last line uses the fact that $R$, $S$, and $W$ are independent in $\mathbf{P}_A[\cdot]$ [Def. 1.36]. This rearranges to

$$\begin{aligned}
\mathbf{E}_A[W] &= \frac{\mathbf{E}[(R - S)^2] - \mathbf{E}_A\Big[\big((R - S - W)^+\big)^2\Big]}{2\,\mathbf{E}[R - S]} \\
&= \frac{\mathbf{Var}[R - S]}{2\,\mathbf{E}[R - S]} + \frac{\mathbf{E}[R - S]^2 - \mathbf{E}_A\Big[\big((R - S - W)^+\big)^2\Big]}{2\,\mathbf{E}[R - S]}.
\end{aligned} \tag{1.9}$$

The second on the right-hand side of (1.9) is related to the length distribution of *idle periods* of the G/G/1, namely the distribution of how long periods of zero work are

[Exr. 1.20]. This is a famously intractable quantity to analyze exactly, but we can still bound it. First, by Jensen's inequality,

$$\mathbf{E}_A\left[\left((R-S-W)^+\right)^2\right] \geq \mathbf{E}_A[(R-S-W)^+]^2. \tag{1.10}$$

This means it suffices to understand just $\mathbf{E}_A[(R-S-W)^+]$, which is significantly easier. Choosing $f(x) = x$ in (1.8) yields

$$\mathbf{E}_A[W] = \mathbf{E}_A[(W+S-R)^+] = \mathbf{E}_A[W] - \mathbf{E}[R-S] + \mathbf{E}_A[(R-S-W)^+],$$

and therefore

$$\mathbf{E}_A[(R-S-W)^+] = \mathbf{E}[R-S]. \tag{1.11}$$

Combining (1.9–1.11) yields a result called the *Kingman bound* [5],

$$\mathbf{E}_A[W] \leq \frac{\mathbf{Var}[R-S]}{2\,\mathbf{E}[R-S]} = \frac{\frac{1}{2}(c_R^2 + \rho^2 c_S^2)}{1-\rho}\,\mathbf{E}[R], \tag{1.12}$$

where $c_R^2$ and $c_S^2$ are the squared coefficients of variation of $R$ and $S$, respectively.

The Kingman bound (1.12) gives us another instance of "work = intensity × variability". Just as job size variance impacts work, which we know from the M/G/1 formula (1.3), we see we get a similar impact from *interarrival time variance*, as captured by the appearance of $\mathbf{Var}[R]$ and $c_R^2$ in (1.12).

*Remark 1.40.* Virtually all of the reasoning above works even if each job size $S$ and the *following* interarrival time $R$ are *not* independent. Specifically, you can check that all of expressions above where $R$ and $S$ appear only as part of an $R-S$ term hold under the weaker assumption that the $(R, S)$ pairs are i.i.d. More formally, the weaker assumption is that the $\left(A_{i+1} - A_i; S(A_i)\right)$ pairs are i.i.d. in $\mathbf{P}_A[\cdot]$.

## 1.5 Exercises

### 1.5.1 Understanding Palm expectation

**Exercise 1.1.**

(a) Show that every *rational* $u > 0$ yields the same value in Definition 1.20. If you prefer, you can restrict to *dyadic* rationals, i.e. $u = n/2^k$ for $n, k \in \mathbb{Z}$. *Hint:* Use linearity of expectation and stationarity.

(b) Assuming $X \geq 0$, extend your argument to irrational $u > 0$. (The conclusion for general $X$ then follows by decomposing $X = X^+ - X^-$.)

*Solution on page 63.*

**Exercise 1.2.** Let $A$ and $B$ be jointly stationary point processes that are (almost surely) disjoint.

(a) Show
$$\lambda_{A\cup B} = \lambda_A + \lambda_B.$$

(b) Show that for any process $X$ that is jointly stationary with $A$ and $B$,
$$\lambda_{A\cup B}\,\mathbf{E}_{A\cup B}[X] = \lambda_A\,\mathbf{E}_A[X] + \lambda_B\,\mathbf{E}_B[X].$$

This result is especially handy when applying RCL 1.22 to unions of point processes.

*Solution on page 64.*

**Exercise 1.3.** Let $A$ be a Poisson process. Explain why for any bounded function $f$,
$$\mathbf{E}[f(A \cup \{t\})] = \mathbf{E}^t_A[f(A)].$$

That is, explain why the distribution of $A \cup \{t\}$ under $\mathbf{P}[\cdot]$ is the same as the distribution of $A$ under $\mathbf{P}^t_A[\cdot]$. *Hint:* How would you generate a sample of the entire Poisson process $A$ "starting at" time $t$?

*Solution on page 65.*

### 1.5.2 Practice with RCL and PIF

**Exercise 1.4.** Let $A$ be a stationary point process. Show
$$\lambda_A = \frac{1}{\mathbf{E}_A[A_1]}.$$

The intuition is that the rate $\lambda_A$ of the point process is the reciprocal of the average amount of time between its points, namely $\mathbf{E}_A[A_1] = \mathbf{E}_A[A_1 - A_0]$.

*Solution on page 66.*

**Exercise 1.5.** Let $W$ be a stationary M/G/1 work process. Show
$$\rho = \mathbf{P}[W > 0].$$

*Hint:* Use RCL 1.22, remembering the rule of thumb: consider, roughly, the integral of the function you're finding the expectation of. Thinking of $\mathbf{P}[W > 0] = \mathbf{E}[\mathbb{1}(W > 0)]$ as an expectation of a "zeroth-order" function of $W$, what does the rule of thumb suggest?

*Solution on page 67.*

**Exercise 1.6.** Let $X$ and $A$ be jointly stationary.

(a) Show
$$\mathbf{E}_A[X(A_1)] = \mathbf{E}_A[X(A_0)].$$

*Hint:* The right-hand side can be more simply written as $\mathbf{E}_A[X]$, because $A_0 = 0$ under $\mathbf{P}_A[\cdot]$ [Def. 1.20]. It's written the way it is as a suggestion of how you might use RCL 1.22 to prove it.

(b) Show that for all $i \in \mathbb{Z}$,

$$\mathbf{E}_A[X(A_i)] = \mathbf{E}_A[X(A_0)].$$

*Hint:* You can either adapt the argument you used for (a), or you can directly apply the result of (a) to $t \mapsto X(\text{something with } t)$.

*Solution on page 68.*

**Exercise 1.7.** In this problem, you will explore variants of PIF 1.25. Let $X \geq 0$ and $A$ be jointly stationary.

(a) Show

$$\mathbf{E}[X] = \lambda_A \, \mathbf{E}_A\left[\int_{A_{-1}}^{0} X(u)\,\mathrm{d}u\right].$$

(b) Show

$$\mathbf{E}[X] = \lambda_A \, \mathbf{E}_A\left[\int_{\frac{1}{2}A_{-1}}^{\frac{1}{2}A_1} X(u)\,\mathrm{d}u\right].$$

*Solution on page 69.*

**Exercise 1.8.** Let $W$ be a stationary M/G/1 work process. Let a *(maximal) busy period* be a maximal contiguous interval of times $t$ during which $W(t) > 0$. Let $B$ and $C$ be the ends and starts, respectively, of busy periods.[7] You may take as given that $B$ and $C$ are jointly stationary with $W$.

(a) Find $\lambda_C$, the average rate with which busy periods start. *Hint:* Think about the relationship between $C$ and the arrival times point process, then use PASTA 1.27.

(b) Find $\mathbf{E}_C[B_1]$, the mean length of a busy period. *Hint:* Use PIF 1.25.

*Hint:* You might find previous exercise helpful for both parts.

*Solution on page 71.*

### 1.5.3 Increasingly unwieldy PASTA

**Exercise 1.9** (Cavatappi). In this problem, you will prove an easy special case of PASTA 1.27 using PIF 1.25. Let $X \geq 0$ and $A$, a Poisson process, be jointly stationary and *independent*. Show

$$\mathbf{E}[X] = \mathbf{E}_A[X].$$

Specifically, use the fact that $A_1 \sim \mathrm{Exp}(\lambda_A)$ under $\mathbf{P}_A[\cdot]$ [Def. 1.14] to show

$$\lambda_A \, \mathbf{E}_A\left[\int_{0}^{A_1} X(u)\,\mathrm{d}u\right] = \mathbf{E}_A[X(A_1)],$$

[7] Formally, we could define $B$ and $C$ as the processes satisfying the following mutual recursion: $B_1(C_0(t)) = \min\{u > C_0(t) : W(u) = 0\}$, and $C_1(B_0(t)) = \min\{u > B_0(t) : W(u+) > 0\}$. But for this problem, you should manage with the "ends and starts of busy periods" definition.

then conclude using Exercise 1.6. *Hint:* It often helps to rewrite a random-domain integral as a deterministic-domain integral with an indicator in the integrand. Also, if $A$ and $X$ are independent, then $A$ and $\mathrm{J}X$ are (almost surely) disjoint, so you can assume $X(a-) = X(a+)$ for all $a \in A$.

*Solution on page 72.*

**Exercise 1.10** (Fettuccine). In this problem, you will prove a relatively easy, but still very useful, special case of PASTA 1.27 for the M/G/1. Let $W$ be a stationary M/G/1 work process, and let $f$ be a nonnegative function. Following the approach from Exercise 1.9, prove

$$\mathbf{E}[f(W)] = \mathbf{E}_A[f(W)],$$

where we recall that $W$ is left-continuous, so $W = W_{(-)}$ [Def. 1.13(c)]. You may take as given the fact that $W_{(+)}$ and $A_1$ are independent under $\mathbf{P}_A[\cdot]$. *Hint:* However, this does not imply that $W(u)$ is independent of $A_1$ under $\mathbf{P}_A[\cdot]$ for $u > 0$. Either argue why this extra independence holds, or come up with an approach that only needs independence of $W_{(+)} = W(0+)$ and $A_1$.

*Solution on page 73.*

**Exercise 1.11** (Capellini). Prove PASTA 1.27. You may use the fact that if $A$ is a Poisson process, then as $\delta \longrightarrow 0+$,

$$\mathbf{P}[N_A(0,\delta) = 1] \sim \lambda_A \delta,$$
$$\mathbf{P}[N_A(0,\delta) \geq 2] \leq O(\delta^2).$$

(a) Give a proof under the assumption that $X$ is continuous and $|\mathrm{D}X|$ is bounded.

(b) *Challenge!* If you really like real analysis, try to extend (a) to a less stringent condition.

### 1.5.4 Increasingly fancy M/G/1 variants

**Exercise 1.12.** Let $W$ be a stationary M/G/1 work process. Find a formula for $\mathbf{E}[e^{\theta W(t)}]$ using RCL 1.22 and PASTA 1.27.

(a) Do this assuming $\theta \leq 0$. *Hint:* What, roughly, is the integral of $w \mapsto e^{\theta w}$?

(b) *Challenge!* Do this assuming $\theta > 0$, obtaining the same formula as in (a), but carefully tracking what preconditions are needed to ensure $\mathbf{E}[e^{\theta W(t)}] < \infty$. *Hint:* Apply RCL 1.22 to a truncated version of what you used in (a). You might get quantities that you can't analyze exactly, but you can bound them.

(c) *Open-ended….* Can you find the secret buses [§ 1.4.2] hiding your formula?

*Solution on page 74.*

**Exercise 1.13.** The *M/G/1 with setup times (M/G/1/setup)* is a variant of the M/G/1, but with the following change: whenever a job arrives to an empty system, in addition to the job's size $S$ being added to the work, an additional *setup time*, sampled i.i.d. from a distribution $U$ on $[0, \infty)$, is also added. This represents the server taking extra time $U$ to set up after being idle.

Let $W$ be a stationary standard M/G/1/setup work process, and let $W_{\mathrm{M/G/1}}$ be a standard M/G/1 work process with the same arrival rate and size distribution.

(a) Find a formula for $\mathbf{E}[W]$ of the form

$$\mathbf{E}[W] = \mathbf{E}[W_{\mathrm{M/G/1}}] + \text{something},$$

where the M/G/1 and M/G/1/setup have the same arrival rate and size distribution.

(b) Can you interpret the "something" in your answer to (a) as the mean of some distribution? What might that distribution represent?

(c) Find a formula for $\mathbf{E}[e^{\theta W}]$. You should find a similar decomposition to what you found in (a). You may assume $\theta \leq 0$.

(d) Based on your answer to (c), was the distribution you found in (b) was correct, or did you find a different distribution that happens to have the right mean?

*Solution on page 76.*

**Exercise 1.14.** Repeat Exercise 1.13, but for the *M/G/1 with vacations (M/G/1/vacation)*. In the M/G/1/vacation, whenever the work reaches 0, it immediately jumps up by a *vacation* amount, which is sampled i.i.d from a distribution $V$ on $(0, \infty)$. This represents server taking a break whenever there's no work to do, coming back after time $V$.

The main difficulty of this problem is that in the M/G/1/vacation, $\Delta W$ is not simply the arrival times $A$. But you can partition $\Delta W = A \cup B$, where $B$ is the times when vacations start. One can show that $A$ and $B$ are (almost surely) disjoint, and you may use this fact without proof for this problem. *Hint:* You might find Exercise 1.2 handy. You can't use PASTA 1.27 under $\mathbf{P}_B[\cdot]$, so hopefully you won't need to....

*Solution on page 79.*

**Exercise 1.15.** *Challenge!* The *M/G/k* is a *multiserver* variant of the M/G/1. Specifically, let's imagine that the M/G/$k$ has $k$ "slow" servers, which run $k$ times slower than the single server of the M/G/1. A job of size $s$ thus takes time $ks$ to finish on one of the slow servers, but this is balanced out by the fact that there are $k$ servers. If all $k$ servers are busy at time $t$, then the M/G/$k$ is still completing work at rate 1, so $\mathrm{D}W(t) = -1$.

However, the M/G/$k$'s work process $W$ is *not standard* [Def. 1.13], because if there are fewer than $k$ jobs in the system at time $t$, then $\mathrm{D}W(t) \neq -1$, even if $W(t) > 0$. In fact, this reveals that $\mathrm{D}W(t)$ is no longer a deterministic function of $W(t)$: it depends on the number of jobs in the system, which we can't infer from $W(t)$ alone.

The above difficulties make analyzing the M/G/$k$'s mean work $\mathbf{E}[W]$ intractable in general. However, we can still get some useful formulas which lead to bounds under some conditions. The key idea is to define an *idleness process* [Def. 2.18]

$$I(t) := 1 - \frac{\text{\# jobs present at } t}{k} = \text{fraction of servers that are idle at } t,$$

then express $\mathrm{D}W(t)$ in terms of $I(t)$. You may assume $W$ and $I$ are jointly stationary.

(a) Show

$$\mathbf{E}[W] = \mathbf{E}[W_{\mathrm{M/G/1}}] + \frac{\mathbf{E}[IW]}{1-\rho}.$$

(b) Assuming there exists $m$ such that $\mathbf{P}[S \le m] = 1$, show

$$\mathbf{E}[W] \le \mathbf{E}[W_{\mathrm{M/G/1}}] + (k-1)m.$$

*Hint:* If $I(t) > 0$, how many jobs can there possibly be in the system at time $t$?

(c) Try to give an intuitive interpretation of the $\mathbf{E}[IW]/(1-\rho)$ term. *Hint:* Here's one somewhat heavy approach. Define the Palm-like expectation $\mathbf{E}_I[X] = \mathbf{E}[IX]/(1-\rho)$ for $X$ jointly stationary with $W$ and $I$. Just as $\mathbf{E}_A[\cdot]$ captures the perspective of an arriving job, consider: what perspective does $\mathbf{E}_I[\cdot]$ capture?

(d) Find a formula for $\mathbf{E}[e^{\theta W}]$ analogous to (a). You may assume $\theta \le 0$. *Hint:* You should get $\mathbf{E}[e^{\theta W_{\mathrm{M/G/1}}}]$ [Exr. 1.12] times a factor that can be written using $\mathbf{E}_I[\cdot]$.

(e) *Open-ended....* To what extent do the above results generalize beyond the M/G/$k$?

*Solution on page 81.*

### 1.5.5 Buses and recurrence time

**Exercise 1.16.** Let a *$(\lambda, S)$-bus process* be, roughly, an $S$-bus process [Def. 1.31] where buses mercifully pause at the bus stop for $\mathrm{Exp}(\lambda)$ time. More formally, let arrivals $A$ and departures $B$ be point processes defined recursively by fresh sampling:

$$\begin{aligned} A_i - B_{i-1} &\sim S, \\ B_i - A_i &\sim \mathrm{Exp}(\lambda). \end{aligned}$$

Then the standard work (or "wait") process

$$W(t) := \begin{cases} A_1(t) - t & \text{if } A_1(t) < B_1(t) \\ 0 & \text{otherwise} \end{cases}$$

is a $(\lambda, S)$-bus process.

Provided $\mathbf{E}[S] < \infty$ and $\lambda \in (0, \infty)$, one can show that a stationary $(\lambda, S)$-bus process $W$ exists, so we restrict attention to this case. Note that if $W$ is stationary, then it is jointly stationary with $A$ and $B$, because $A$ and $B$ can be expressed as functions of $W$. *Hint:* Despite the $\mathrm{Exp}(\lambda)$, neither $A$ nor $B$ is a Poisson process, so you can't use PASTA 1.27.

In this problem, you will adapt Theorem 1.32 to $(\lambda, S)$-bus processes, then compare your results. Specifically:

(a) Find $\mathbf{E}[W]$. You may assume $\mathbf{E}[S^2] < \infty$.

(b) Find $\mathbf{E}[e^{\theta W}]$. You may assume $\theta \le 0$.

(c) Find $\mathbf{P}[W > x]$.

(d) Describe $W$'s distribution in terms of the excess distribution $S_{\mathrm{e}}$ [Def. 1.33].

(e) Describe how "work ≈ intensity × variability" manifests in your $\mathbf{E}[W]$ formula.

(f) *Open-ended....* What is the most general version of this result you can state and prove? Can $\mathrm{Exp}(\lambda)$ be replaced by an arbitrary distribution $Z > 0$? Can $A$, $B$, or both be replaced with an arbitrary stationary point processes?

**Exercise 1.17.** Let $A$ be an $S$-renewal process, let $W := A_1(t) - t$ be its bus process, and let its *reverse-bus process* be

$$V(t) := t - A_0(t),$$

i.e. the time since the previous bus arrived. Find $\mathbf{P}[V > x; W > y]$. *Hint:* Use PIF 1.25.

The significance of this problem is that, as mentioned in Remark 1.35, it tells us how to set the joint distribution of $A_0$ and $A_1$ such that $A$ is stationary. This is because at $t = 0$, we have $V = -A_0$ and $W = A_1$.

**Exercise 1.18.** *Challenge!* In this problem, we will give a more rigorous account of the "work layers" intuition of the M/G/1 introduced in Section 1.4.2. As discussed there, we will assume PLCFS scheduling, so that each layer corresponds to a job. You will show that

- the number of jobs $N$ has distribution $\mathrm{Geo}_0(1-\rho)$; and
- conditional on $N$, the $N$ jobs' remaining work amounts $L_1, \dots, L_N$ are i.i.d. with distribution $S_{\mathrm{e}}$.

You should feel comfortable with Exercise 1.12(a) before attempting this problem.

Before we state the result you will show, let's be a little more precise about what $N$ and $L_i$ are as processes. Let[8]

$$N(t) := \text{number of jobs present at time } t,$$
$$L_i(t) := \begin{cases} \text{remaining work of } i\text{th most recently arrived job} & \text{if } N(t) \ge i \\ 0 & \text{if } N(t) < i. \end{cases}$$

We assume the sequence-valued process $(L_1, L_2, \dots)$ is stationary. This means it is jointly stationary with $W$ and $N$, because they can be expressed as

$$W = \sum_{i=1}^{\infty} L_i, \qquad N = \max\{i \in \mathbb{N} : L_i > 0\}.$$

[8] The subscripts on the $L_i$ below don't quite match with the subscripts on the $L_i$ in Figure 1.3. Below, the subscripts of layers in the system are always $1, \dots, N$. But in Figure 1.3, a layer's subscript corresponds to a job's *absolute* arrival index, so the subscripts of layers in the system need not be contiguous.

Your will find the *joint transform* of the $L_1, L_2, \dots$ sequence. Specifically, let $\theta_1, \theta_2, \dots$ be a real sequence, which you may assume is nonpositive, and let

$$X_k := \prod_{i=1}^{k} e^{\theta_i L_i} = \exp\left(\sum_{i=1}^{k} \theta_i L_i\right).$$

Your task is to show that for all $k \in \mathbb{N}$,

$$\mathbf{E}[X_k \, \mathbb{1}(N = k)] = (1-\rho)\rho^k \prod_{i=1}^{k} \mathbf{E}[e^{\theta_i S_\mathrm{e}}], \tag{1.13}$$

then explain how this yields the claims at the start of this problem.

(a) Show (1.13) for $k = 0$.

(b) Use RCL 1.22 to express $\mathbf{E}[X_k \, \mathbb{1}(N = k)]$ in terms of $\mathbf{E}[X_{k-1} \, \mathbb{1}(N = k-1)]$, then show (1.13) for all $k$ by induction. *Hint:* You might be tempted to put indicators in the process you apply RCL 1.22 to. This can work, but it's a bit of a trap: you'll likely get jumps when jobs *depart*, which we don't yet have great tools for handling. Try to choose a process that only jumps when jobs *arrive*, which makes PASTA 1.27 applicable. If in doubt, try applying RCL 1.22 to $X_k$, then see what happens.

(c) Use (1.13) to show $N \sim \mathrm{Geo}_0(1-\rho)$ by finding $\mathbf{P}[N = k]$.

(d) Combine (1.13) and (c) to conclude that the conditional joint transform of $L_1, \dots, L_k$ given $N = k$, namely $\mathbf{E}[X_k \mid N = k]$, is the joint transform of $k$ i.i.d. samples from $S_\mathrm{e}$.

(e) Use (1.13) to obtain a formula for $\mathbf{E}[e^{\theta W}]$, and check that it agrees with your answer to Exercise 1.12.

(f) *Open-ended....* Try adapting (1.13) to the M/G/1/setup [Exr. 1.13] or M/G/1/vacation [Exr. 1.14]. You'll have to think about how to define the "work layers" corresponding to setups or vacations.

### 1.5.6 Non-Poisson arrivals

**Exercise 1.19.** Let $W$ be a stationary G/G/1 work process.

(a) Express $\mathbf{E}[W]$ in terms of $\mathbf{E}_A[W]$. *Hint:* Either RCL 1.22 or PIF 1.25 can work

(b) Use (a) and material from Section 1.4.3 to upper bound $\mathbf{E}[W]$ similarly to (1.12).

(c) *Without* using material from Section 1.4.3, find a second relationship between $\mathbf{E}[W]$ and $\mathbf{E}_A[W]$, then combine it with (a) to prove (1.9). *Hint:* Try whichever of RCL 1.22 or PIF 1.25 you didn't use in (a).

**Exercise 1.20.** *Challenge!* Let $W$ be a stationary G/G/1 work process. Let a *(maximal) idle period* be a maximal contiguous interval of times $t$ such that $W(t) = 0$. Let $U$ be the distribution of the length of idle periods. Specifically, letting $B$ be the stationary point process of the starts of idle periods, let $U$ be the distribution of $A_1$ in $\mathbf{P}_B[\cdot]$.

Express $\mathbf{E}_A\left[\left((R-S-W)^+\right)^2\right]$, the intractable term from the G/G/1 $\mathbf{E}_A[W]$ formula (1.9), in terms of the excess $U_\mathrm{e}$.

**Exercise 1.21.** *Open-ended….* Let $W$ be a stationary G/G/1 work process. Try to adapt the strategy used in Section 1.4.3 to find a formula for the work transform $\mathbf{E}_A[e^{\theta W}]$. Make whatever additional assumptions you deem necessary. Can you get an exact formula that contains some intractable terms like (1.9)? How about a bound like (1.12)?

**Exercise 1.22.** *Open-ended….* Consider a system where jobs are scheduled to arrive with deterministic gaps between arrivals of $1/\lambda$ seconds, but, relatably, jobs always arrive late. Specifically, suppose a job that is scheduled to arrive at time $t$ actually arrives at time freshly sampled from $\mathrm{Unif}(t, t + \alpha/\lambda)$, where $\alpha \in [0, 1]$.

(a) Formalize the above description into a definition of an arrival time point process $A$.

(b) Let $W$ be the standard work process whose jumps are the $S$-GD/G arrival process of $A$. Try to bound $\mathbf{E}[W]$ from both above and below. Feel free to prove a bound that relies on finiteness for whatever expectations related to $S$ you need.

(c) Describe how "work ≈ intensity × variability" manifests in your $\mathbf{E}[W]$ bounds.

(d) Repeat (b), but for $\mathbf{E}[e^{\theta W}]$. *Hint:* To avoid trivial bounds, e.g. $0 \le \mathbf{E}[e^{\theta W}] \le 1$ for $\theta \le 0$, try to prove bounds that you believe hold for $\theta > 0$, even if you don't work through all the details as rigorously as in Exercise 1.12(b).

(e) *Challenge!* Can you extend any of your bounds to $\alpha \in (1, 2]$? How about $\alpha > 2$?

# Chapter 2
# Dispatching and state-space collapse

Having spent Chapter 1 developing tools to *analyze* work in queues, we're now ready to solve our first *control* problems in queues. The specific type of control we'll consider is *dispatching decisions*. In the dispatching system we'll study in this chapter, whenever a job arrives, we must immediately send it to one of multiple queues, as shown in Figure 2.1. Our objective will be to minimize the *mean response time*, where a job's response time is the total amount of time it spends in the system between arrival and departure.

It turns out that even the very simplest dispatching problems are usually intractable to solve exactly! As such, we'll have to settle for near-optimal policies. To narrow our focus, we'll consider a dispatching system with M/G arrivals [Def. 1.15], and we'll focus on achieving near-optimality in *heavy traffic*, meaning when the load is near the maximum. Even this problem was solved only recently [9].

The main technical tool that will enable us to analyze dispatching policies is *state-space collapse*. Roughly speaking, state-space collapse occurs when a system's state nearly always stays in, or at least nearby, a subset of the possible states. For example, we will soon encounter a dispatching policy called *Least Work Left (LWL)* that tries to keep the amount of work balanced across two queues. Modeling the state of this system as the pair $(W_1, W_2)$, where $W_i$ is the amount of work in queue $i$ [Fig. 2.1], the full state space is $[0, \infty)^2$, but LWL induces state-space collapse to the subset where $W_1 = W_2$. Beyond LWL, we'll see that optimal dispatching largely boils down to inducing the right type of state-space collapse.

## 2.1 Setting: the heavy-traffic M/G/2/dispatch

We consider a dispatching system with two queues, as pictured in Figure 2.1. We write $W_1$ and $W_2$ for the work processes of the two queues. We assume an M/G arrival process with arrival rate $\lambda$, job size distribution $S$, and load $\rho$. This creates a system which we call the *M/G/2/dispatch*, defined formally in Definition 2.3 below.

### 2.1.1 Marks on point processes

Below, we will introduce notation for the queue to which a job is dispatched [Def. 2.3(a)] and the amount of time a job spends in the system [Def. 2.7]. These concepts really only make sense when we're talking about a specific job. We'll integrate this into our notation as a process that we only need to define at arrival times, which we call a *mark process* on the arrival times.

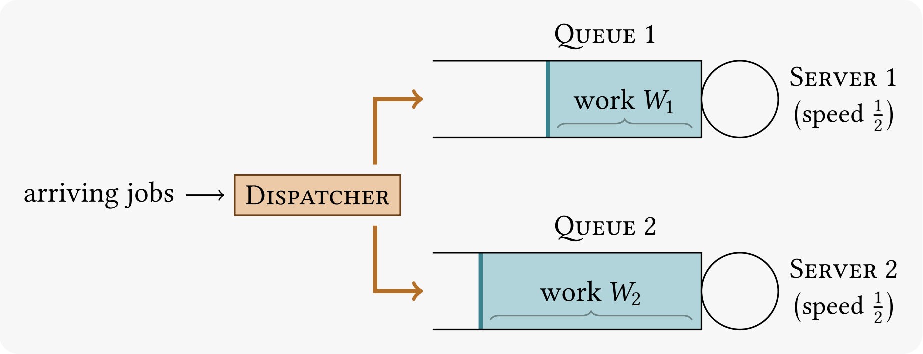


**Figure 2.1.** A two-server dispatching system. Whenever a job arrives, the dispatcher immediately sends it to one of the two servers, where it waits in a server-specific queue. The work waiting in queue $i$ is denoted $W_i$. Both servers have speed $1/2$, so $W_i$ decreases at rate $1/2$ while it is nonzero. This speed convention makes the system comparable to a single-server queue with speed 1.

**Definition 2.1.** Let $A$ be a point process. A *mark process* on $A$ is a right-continuous, piecewise-constant process that only changes value at times in $A$. Equivalently, $M$ is a mark process on $A$ if for all $t$,

$$M(t) = M(A_0(t)).$$

This means that when defining a mark process $M$ on $A$, it suffices to define $M(t)$ only for $t \in A$. We also define

$$M_i(t) \coloneqq M(A_i(t)),$$

which generalizes $M = M_0$.

### 2.1.2 The M/G/$k$/dispatch queueing system

**Definition 2.2.** A stochastic process $W$ is a *speed $c$ standard work process* if $W/c$ is a standard work process [Def. 1.13, Fig. 1.1]. That is, $W$ is like a standard work process, but it decreases at rate $c$ instead of rate 1. We define speed $c$ versions of specific types of standard work processes similarly, e.g. a *speed $c$ M/G/1 work process* [Def. 1.16].

**Definition 2.3.** A *$(\lambda, S)$-M/G/$k$/dispatch* is a queueing system with $k$ servers, each with its own queue. Jobs arrive according to a $(\lambda, S)$-M/G arrival process, and whenever a job arrives, it is immediately sent to one of the $k$ servers.

(a) The *dispatch index* $J$ is a mark process on arrival times $A$, where $J(t)$ is the server to which the arrival at time $t \in A$ is dispatched.

(b) The *queue $i$ work process*, denoted $W_i$, is a speed $1/k$ standard work process with jump

$$\Delta W_i(t) \coloneqq S(t)\, \mathbb{1}(J(t) = i).$$

(c) The *(total) work process* $W$ is
$$W := \sum_{i=1}^{k} W_i.$$

(d) For $k = 2$, the *(work) gap process* $G$ is
$$G := |W_1 - W_2|,$$
which means
$$\min\{W_1, W_2\} = W - G,$$
$$\max\{W_1, W_2\} = W + G.$$

We define dispatching systems with other arrival processes analogously, with names based on Notation 1.39. For instance, the *G/G/k/dispatch* is defined as above, but with G/G arrivals [Def. 1.36] instead of M/G arrivals.

For the most part, we work with the M/G/2/dispatch, though you'll explore the M/G/$k$/dispatch in some of the exercises [Exrs. 2.8, 2.11]. This choice is mostly for ease of presentation: many of the results and takeaways apply for all $k$, but they require more involved notation.

*Remark 2.4.*

(a) The reason we define the M/G/$k$/dispatch to use speed $1/k$ servers instead of speed 1 servers is to make it comparable to the standard M/G/1. In particular, we will see that the total work in an M/G/$k$/dispatch is bounded below by the work in an M/G/1.

(b) If $W$ is a speed $c$ standard work process, then $t \mapsto W(t/c)$ is a speed 1 standard work process with the same stationary distribution, as well as the same distribution of $\Delta W$ under $\mathbf{P}_{JW}[\cdot]$.

(c) In the M/G/$k$/dispatch, with the right parameterization of the arrival process, the stationary distribution of the queue $i$ work processes $W_i$ would be the same if we used a speed 1 convention instead of our speed $1/k$ convention. Specifically, in light of Remark 2.4(b), if we write a formula for $\mathbf{E}[W_i]$ in terms of $S$ and $\rho$ but *not* $\lambda$, it remains valid under either server speed convention.

In Definition 2.3(a), the concrete definition of $J$ is left unspecified. This is because $J$ depends on the *dispatching policy*, meaning the rule that decides which server to send each arrival to. The main aim of this chapter is to develop tools for designing and analyzing dispatching policies.

*Notation 2.5.*

(a) We usually denote a generic policy by the variable $\pi$, and we denote specific policies by their (possibly abbreviated) name.

(b) We write policies in the subscript when helpful for disambiguation, e.g. $W_\pi$ or $J_\pi$. But for the most part, we focus on analyzing one policy at a time, in which case we usually omit the subscript.

(c) We also sometimes use a subscript “M/G/1” to denote a quantity in a standard M/G/1 [Def. 1.16], using first-come first-served [§ 2.1.3] if the scheduling policy is relevant.

**Assumption 2.6.** We make the following assumptions in this chapter.

(a) Unless specified otherwise, we assume that all of the M/G/$k$/dispatch-related processes are *jointly stationary*. This amounts to assuming the existence of $J, W_1, \dots, W_k$ that are jointly stationary with the M/G arrival process, as everything else we define will be in terms of those.

(b) We assume $\mathbf{E}[S^2] < \infty$. This is essential assumption, because if it doesn't hold, then $\mathbf{E}[W_{\mathrm{M/G/1}}] = \infty$ [(1.3)], which leads to all of the policies we study having infinite mean response time in heavy traffic [§ 2.1.3].

(c) For the most part, we assume $S$ is continuous. This is only for ease of presentation, as the main obstacle has a straightforward workaround [Rmk. 2.16].

(d) Even when we consider non-continuous $S$, we still assume $S$ is not deterministic. This enables us make strict some inequalities [(2.4), (2.5)] that are equalities only in the deterministic case.

### 2.1.3 Metric: heavy-traffic mean response time

**Definition 2.7.**

(a) A job's *waiting time* $T^{\text{wait}}$ is the total amount of time it spends in the system prior to the first moment it enters service.

(b) A job's *residence time* $T^{\text{resd}}$ is the total amount of time it spends in the system after it first enters service.

(c) A job's *response time* $T = T^{\text{wait}} + T^{\text{resd}}$ is the total amount of time it spends in the system.

Formally, if $A$ is the arrival times point process, we let $T$ be a mark process on $A$, with $T(t)$ denoting the response time of the job that enters the system at time $t \in A$, and similarly for $T^{\text{wait}}$ and $T^{\text{resd}}$.

Of course, a job's response time $T$ depends on the dispatching policy, just like the dispatch index $J$. In general, response time can also depend on the scheduling policy in use at each queue. In this chapter, we'll assume that each queue uses *First-Come First-Served (FCFS)*. This means a job's waiting time $T^{\text{wait}}$ is simply the server speed times the work $W_J$ at the server it is dispatched to, so in the M/G/$k$/dispatch, for $t \in A$,

$$T^{\text{wait}}(t) = kW_J(t), \qquad T^{\text{resd}}(t) = kS(t), \qquad T(t) = kW_J(t) + kS(t). \tag{2.1}$$

Our aim is to design a dispatching policy $\pi$ for the M/G/2/dispatch that minimizes mean response time $\mathbf{E}_A[T_\pi]$. Thanks to (2.1), this amounts to minimizing $\mathbf{E}_A[W_J]$. We call

$W_J$ the *observed work*, because it's the amount of work observed by an arrival at the server to which it is dispatched, excluding its own size [Def. 1.13(c)].

However, it turns out that minimizing $\mathbf{E}_A[T]$ in the M/G/2/dispatch is essentially intractable to solve exactly. As such, we'll focus on the *heavy-traffic* regime, defined below, where the problem is theoretically solvable. The hope is that solving the problem in heavy traffic teaches us lessons that are relevant even outside of heavy traffic.

**Definition 2.8.** When working with M/G arrivals, the *heavy-traffic limit* refers to the limit as $\lambda \to 1/\mathbf{E}[S]$, with $S$ remaining fixed. We write this as $\rho \to 1$ throughout. In informal discussion, phrases like "heavy-traffic" and "in heavy traffic" refer to this $\rho \to 1$ limit.

One way to view the $\rho \to 1$ notation is to view $\rho$ and $S$ as the defining parameters of the M/G arrival process, from which the arrival rate is defined as $\lambda \coloneqq \rho/\mathbf{E}[S]$.

*Notation 2.9.* We use the following notation for asymptotic comparisons.[1] Suppose we are considering positive functions $f(x), g(x) > 0$ of $x$ in an $x \to y$ limit.

(a) $f(x) \ll g(x)$ means $\lim_{x\to y} f(x)/g(x) = 0$.

(b) $f(x) \gg g(x)$ means $\lim_{x\to y} f(x)/g(x) = \infty$.

(c) $f(x) \lesssim g(x)$ means $\limsup_{x\to y} f(x)/g(x) < \infty$.

(d) $f(x) \sim g(x)$ means $\limsup_{x\to y} f(x)/g(x) < \infty$ and $\liminf_{x\to y} f(x)/g(x) > 0$.

(e) $f(x) \gtrsim g(x)$ means $\liminf_{x\to y} f(x)/g(x) > 0$.

(f) $f(x) \lessapprox g(x)$ means $\limsup_{x\to y} f(x)/g(x) \le 1$.

(g) $f(x) \approx g(x)$ means $\lim_{x\to y} f(x)/g(x) = 1$.

(h) $f(x) \gtrapprox g(x)$ means $\liminf_{x\to y} f(x)/g(x) \ge 1$.

*Notation 2.10.* In this chapter, we use the relations in Notation 2.9 only for the $\rho \to 1$ limit, so we often omit the "as $\rho \to 1$" for brevity.

How does mean response time behave in heavy traffic? We know that in an M/G/1 using FCFS, we have

$$\mathbf{E}_A[T_{\mathrm{M/G/1}}] = \mathbf{E}[W_{\mathrm{M/G/1}}] + \mathbf{E}[S] \approx \frac{\mathbf{E}[S_{\mathrm{e}}]}{1-\rho}.$$

It's natural to expect the same $\frac{1}{1-\rho}$ scaling of mean response time in the M/G/2/dispatch. We therefore define the *heavy-traffic constant* of $\pi$ to be

$$C_\pi \coloneqq \lim_{\rho\to 1} (1-\rho)\,\mathbf{E}_A[T_\pi].$$

[1] There's no standard binary relation notation that accounts for all of these. Our choice here is a compromise between existing standards while maintaining internal consistency, e.g. a single $\sim$ squiggle meaning "modulo a multiplicative constant" and a double $\approx$ meaning "with a tight multiplicative constant". It will always be clear from context when $\sim$ means "has distribution".

That is, if $C_\pi \in (0, \infty)$, then

$$\mathbf{E}_A[T_\pi] \approx \frac{C_\pi}{1-\rho}.$$

For example, $C_{\mathrm{M/G/1}} = \mathbf{E}[S_\mathrm{e}]$. We now have the terminology to precisely state the problem we'll solve in the rest of this chapter.

*Question 2.11.* Consider dispatching policies for the M/G/2/dispatch.

(a) What is the optimal heavy-traffic constant $C_* := \inf_\pi C_\pi$?

(b) What dispatching policy $\pi$, if any, achieves tail constant $C_*$?

*Notation 2.12.* By "policy", we often really mean *parameterized policy family.* For example, the dispatching policies we design may depend on the size distribution $S$ and load $\rho$.

## 2.2 Two dispatching policies

How do we achieve low mean response time in the M/G/2/dispatch? There are at least two natural ideas that come to mind, each of which leads to a policy.

- We could dispatch each job in the way that is best for its individual response time. This idea leads to a policy called *Least Work Left (LWL)* [§ 2.2.1].
- We could protect small jobs from waiting behind large jobs by reserving one of the queues for them, much like the "$n$ items or less" checkout lane in a grocery store. This idea leads to a policy called *Size Interval Task Assignment (SITA)* [§ 2.2.2], introduced by [3].

Below, we'll define LWL and SITA more precisely, then compare how they perform [§ 2.2.3].

### 2.2.1 LWL: Least Work Left

**Policy 2.13.** *Least Work Left (LWL)* is the dispatching policy that sends each arriving job to the queue with the least work. That is, under LWL, the dispatch index at arrival times $t \in A$ is

$$J_{\mathrm{LWL}}(t) := \arg\min_i W_i(t),$$

where $W_i(t) = W_i(t-)$ is the work *before* the dispatching occurs [Def. 1.13(c)].

LWL's strategy of sending jobs to the queue with least work has two potential benefits.

- Given queue work amounts $W_1, \dots, W_k$, LWL achieves the minimum possible observed work $W_J = \min_i W_i$.
- By always choosing smallest work amount to increase, LWL tends to balance the work across the $k$ queues evenly.

So, what heavy-traffic constant $C_{\mathrm{LWL}}$ do these properties lead to? It turns out LWL *roughly mimics a single-server M/G/1* for the following reasons.

- Because $W_1$ and $W_2$ are roughly balanced, the total work process $W$ behaves roughly like a standard work process, so

$$\mathbf{E}[W_{\mathrm{LWL}}] \approx \mathbf{E}[W_{\mathrm{M/G/1}}].$$

- Again because $W_1$ and $W_2$ are roughly balanced, the observed work $W_J$ is roughly half of the total work $W$, so

$$\mathbf{E}_A[T_{\mathrm{LWL}}^{\mathrm{wait}}] = 2\,\mathbf{E}[W_J] \approx \mathbf{E}[W_{\mathrm{LWL}}].$$

Combining these observations, we get

$$\mathbf{E}_A[T_{\mathrm{LWL}}] \approx \mathbf{E}[W_{\mathrm{LWL}}] + 2\,\mathbf{E}[S] \approx \mathbf{E}[W_{\mathrm{M/G/1}}], \tag{2.2}$$

which means

$$C_{\mathrm{LWL}} = C_{\mathrm{M/G/1}} = \mathbf{E}[S_{\mathrm{e}}]. \tag{2.3}$$

In fact, this holds for the M/G/$k$/dispatch for general $k \geq 2$ [Exr. 2.11].

We'll formally state and prove the observations above in Section 2.3. For exposition purposes, we'll do so under an extra assumption on $S$, but it can be removed [Exr. 2.10].

### 2.2.2 SITA: Size Interval Task Assignment

**Definition 2.14.**

(a) For a given size distribution $S$ and integer $k \geq 2$, the *load-equalizing thresholds*, denoted $q_1, \dots, q_{k-1}$, are the thresholds such that for all $i \in \{1, \dots, k\}$,

$$\mathbf{E}[S\,\mathbb{1}(S \in [q_{i-1}, q_i))] = \frac{\mathbf{E}[S]}{k},$$

where $q_0 = 0$ and $q_k = \infty$ as edge cases. Such thresholds exist as long as $S$ is a continuous distribution [Asm. 2.6(c)].

(b) When $k = 2$, we write $q \coloneqq q_1$ and call $q$ the *load median*. It is the threshold such that

$$\mathbf{E}[S\,\mathbb{1}(S \leq q)] = \mathbf{E}[S\,\mathbb{1}(S > q)] = \frac{\mathbf{E}[S]}{2}, \tag{2.4}$$

This means

$$\mathbf{P}[S \leq q] = \frac{\mathbf{E}[S]}{2\,\mathbf{E}[S \mid S \leq q]} > \frac{1}{2} > \frac{\mathbf{E}[S]}{2\,\mathbf{E}[S \mid S > q]} = \mathbf{P}[S > q]. \tag{2.5}$$

**Policy 2.15.** The *Size Interval Task Assignment (SITA)* dispatching policy with *thresholds* $r_1, \dots, r_{k-1}$ is the $k$-queue dispatching policy that sends jobs of size $[r_{i-1}, r_i)$ to server $i$, where $r_0 := 0$ and $r_k := \infty$ as edge cases. That is, under SITA, the dispatch index at arrival times $t \in A$ is

$$J_{\mathrm{SITA}}(t) := \begin{cases} 1 & \text{if } S(t) < r_1 \\ 2 & \text{if } S(t) \in [r_1, r_2) \\ & \vdots \\ k-1 & \text{if } S(t) \in [r_{k-2}, r_{k-1}) \\ k & \text{if } S(t) \geq r_{k-1}. \end{cases}$$

When $k = 2$, we write $r := r_1$ for the single threshold.

We usually consider SITA with one of two ways of setting thresholds.

(a) *SITA with load-Equalizing thresholds (SITA-E)* uses the load-equalizing thresholds $r_i = q_i$ [Def. 2.14]. These depend on the size distribution but not the load.

(b) *SITA with Optimized thresholds (SITA-O)* uses whatever thresholds minimize mean response time. These depend on both the size distribution and the load.

*Remark 2.16.* If $S$ is not a continuous distribution, then load-equalizing thresholds might not exist. This is because $x \mapsto \mathbf{E}[S\,\mathbb{1}(S \leq x)]$ is not continuous, and if it jumps over an integer multiple of $\mathbf{E}[S]/k$, then one of the load-equalizing thresholds fails to exist.

Fortunately, this obstacle can be easily overcome using *random tie-breaking*. The intuition is that if we perturbed each job's size by very narrow continuously distributed noise, we would have a continuous size distribution, and thus load-equalizing thresholds. Instead of actually perturbing the sizes, we'll assume that each job arrives with a freshly sampled *tiebreaker* $U \sim \mathrm{Unif}(0,1)$, then compare size-tiebreaker pairs $(s, u)$ lexicographically.

Load-equalizing size-tiebreaker pairs $(q_1, u_1), \dots, (q_k, u_k)$ always exist.

- If $\mathbf{E}[S\,\mathbb{1}(S \leq x)] = \frac{i}{k}\,\mathbf{E}[S]$ for some $x$, then we can set $q_i = x$ and $u \in [0,1]$ arbitrarily.
- If instead $\frac{i}{k}\,\mathbf{E}[S]$ is "skipped" by a jump discontinuity of $x \mapsto \mathbf{E}[S\,\mathbb{1}(S \leq x)]$, then we can set $q_i$ to the value where the relevant jump happens, and set $u_i$ to ensure $\mathbf{E}[S\,\mathbb{1}(S \leq q_i; U \leq u_i)] = \frac{i}{k}\,\mathbf{E}[S]$.

The main benefit of SITA is that small jobs don't have to wait behind large jobs. They aren't even in the same queue! So, what heavy-traffic constant $C_{\mathrm{SITA}}$ does this lead to?

Due to the splitting property of Poisson processes [2, Section 11.7], SITA actually creates *two independent M/G/1 queues*, one with size distribution $(S \mid S < r)$ and the other with size distribution $(S \mid S \geq r)$. This means once we specify the threshold $r$, we can obtain $C_{\mathrm{SITA}}$ using the M/G/1 mean work formula (1.3). We'll discuss SITA-E here, leaving SITA-O for you to think about [Exr. 2.6]. Accounting for the server speed of $1/2$, some computation [Exr. 2.5] yields

$$\mathbf{E}_A[T_{\text{SITA-E}}] = \frac{2\lambda}{1-\rho}\Big(\mathbf{P}[S < q]\,\mathbf{E}[S^2\,\mathbb{1}(S < q)] + \mathbf{P}[S \geq q]\,\mathbf{E}[S^2\,\mathbb{1}(S \geq q)]\Big) + 2\,\mathbf{E}[S], \quad (2.6)$$

and so, because $\lambda \to 1/\mathbf{E}[S]$ as $\rho \to 1$,

$$C_{\text{SITA-E}} = \frac{2}{\mathbf{E}[S]}\Big(\mathbf{P}[S < q]\,\mathbf{E}[S^2\,\mathbb{1}(S < q)] + \mathbf{P}[S \geq q]\,\mathbf{E}[S^2\,\mathbb{1}(S \geq q)]\Big). \tag{2.7}$$

Let's try to unpack $C_{\text{SITA-E}}$. To understand the $\mathbf{E}[S^2\,\mathbb{1}(S < q)]$ terms, we observe using (2.4) and Assumption 2.6(c) that

$$\mathbf{E}[S^2\,\mathbb{1}(S < q)] < q\,\mathbf{E}[S\,\mathbb{1}(S < q)] = q\,\mathbf{E}[S\,\mathbb{1}(S \geq q)] < \mathbf{E}[S^2\,\mathbb{1}(S \geq q)].$$

Along similar lines, (2.5) tells us $\mathbf{P}[S < q] > \frac{1}{2} > \mathbf{P}[S \geq q]$. This means

$$C_{\text{SITA-E}} < \frac{2}{\mathbf{E}[S]}\bigg(\frac{1}{2}\,\mathbf{E}[S^2 \mid S < q] + \frac{1}{2}\,\mathbf{E}[S^2 \mid S \geq q]\bigg) = \frac{\mathbf{E}[S^2]}{\mathbf{E}[S]} = 2\,\mathbf{E}[S_{\mathrm{e}}]. \tag{2.8}$$

Roughly speaking, the more variable $S$, the greater the gap in the inequality is. SITA does well when *balanced load* between the two servers results in *unbalanced arrival rates*, with the smaller-size half of the load consisting of many more jobs than the larger-size half.

### 2.2.3 Which is better: LWL or SITA?

How do LWL and SITA compare in heavy-traffic? Combining (2.3) and (2.8), we see $C_{\text{SITA-E}} < 2C_{\text{LWL}}$, but this is far from conclusive. When is this bound tight enough that LWL is better. When is it loose enough that SITA-E is better? Does using SITA-O instead of SITA-E make a big difference?

To gain some intuition for these questions, Figure 2.2 investigates them for numerically the simplest possible non-deterministic size distribution: a two-point distribution. Specifically, for some $b > a > 0$ and $p \in (0, 1)$, let

$$S := \begin{cases} a & \text{with probability } p \\ b & \text{with probability } 1 - q, \end{cases}$$

so that all jobs are either "small" (size $a$) or "large" (size $b$).

There are two important quantities that determine the relative performance of different dispatching policies.

- The *size ratio* $b/a$ between large and small jobs.
- The *load fraction* made up of each size of job. There are only two sizes, so we can consider just one of them:

  $$\text{size } a \text{ load fraction} := \frac{\mathbf{E}[S\,\mathbb{1}(S = a)]}{\mathbf{E}[S]} = \frac{pa}{pa + (1 - q)b}. \tag{2.9}$$

  If this fraction is greater than $1/2$, then in heavy traffic, SITA sends only small jobs to queue 1 and sends a mix of both sizes to queue 2; and vice versa if it's less than $1/2$ [Rmk. 2.16].

These two quantities determine a policy's heavy-traffic constant up to a normalizing factor, which could be $\mathbf{E}[S]$. In particular, they determine *ratios* of heavy-traffic constants between any two policies.

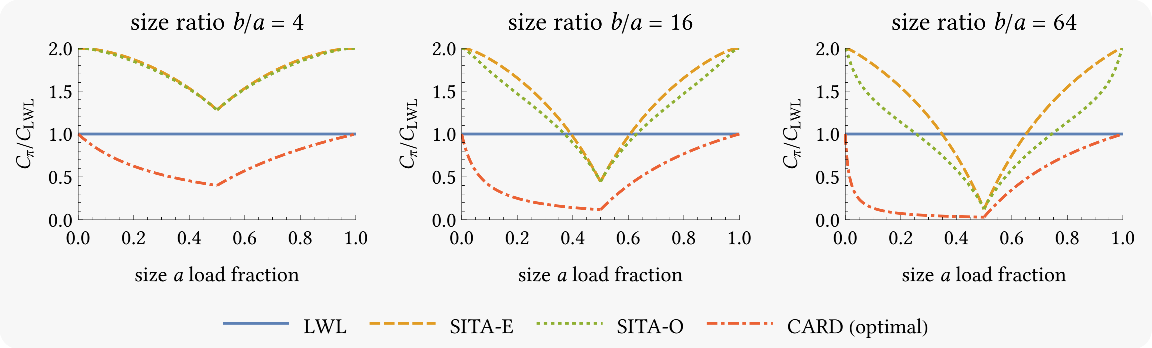


**Figure 2.2.** Normalized heavy-traffic constants $C_\pi/C_{\mathrm{LWL}}$ for a size distribution with two possible sizes $a < b$ ("small" and "large", respectively). The policies are LWL [Pol. 2.13], two variants of SITA [Pol. 2.15], and a policy called CARD introduced later [Pol. 2.32]. The $x$ axis is the fraction (2.9) of load made up of small jobs. SITA beats LWL when the overall size distribution is variable (larger size ratio $b/a$), but the size distribution of the lower and upper halves of the load are less variable (size $a$ load fraction closer to $1/2$). CARD, which minimizes the heavy-traffic constant [§§ 2.4, 2.5], consistently beats both SITA and LWL.

### Explaining the performance differences between LWL and SITA

Figure 2.2 compares the heavy-traffic constants of LWL, SITA-E, and SITA-O as the size ratio and size $a$ load fraction vary. Both SITA variants do best relative to LWL when the size ratio $b/a$ is large, but the load is split roughly equally between the two sizes. Here's one way to understand this intuitively in terms of job size variability [§ 1.4].

- The larger $b/a$ is, the *more variable* the overall size distribution $S$ is, leading to larger $C_{\mathrm{LWL}}$.
- The closer the size $a$ load fraction (2.9) is to $1/2$, the *less variable* the size distributions at each of the two queues is under SITA, leading to smaller $C_{\mathrm{SITA}}$.

However, both of the above factors seem to favor LWL. Why, then, is SITA sometimes much worse than LWL? It's because SITA has, in some sense, a more variable arrival process at each queue.

- SITA uses *open-loop* control, meaning it doesn't use the system state decisions. Specifically, $J_{\mathrm{SITA}}$ doesn't depend on $W_1$ and $W_2$. As explained in Section 2.2.2, SITA results in M/G arrivals at each queue.
- In contrast, LWL uses *closed-loop* control, meaning it uses the system state when making decisions. Specifically, $J_{\mathrm{LWL}}$ is a function of the state $(W_1, W_2)$. As explored in Section 2.3 and Exercise 2.9, LWL effectively "spreads out variability" from the overall M/G arrival process, resulting in less variable arrival processes at each of the two queues. If, for instance, a very large job arrives and is sent to queue 1, LWL compensates by sending the next several arrivals to queue 2.

### Combining the strengths of LWL and SITA

To summarize the above discussion, here are the main strengths of LWL and SITA.

- LWL uses closed-loop dispatching to regulate the arrival processes to the two queues.
- SITA uses size-based dispatching to protect small jobs from getting stuck behind large jobs.

This prompts a natural question: can we combine closed-loop and size-based dispatching to get the best of both worlds? Yes we can! The last policy in Figure 2.2, called CARD [Pol. 2.32], does exactly this, and we see that it outperforms both LWL and SITA, often significantly so.

We'll properly discuss CARD in Section 2.5. For now, here's the high-level idea of how it combines closed-loop and size-based dispatching.

- Like LWL, CARD uses closed-loop control. But instead of keeping $W_1$ and $W_2$ nearly equal, CARD aims for *maximal asymmetry*, keeping $W_1$ much smaller than $W_2$.
- Like SITA, CARD protects small jobs from large jobs. But CARD is much better for the small jobs than SITA: CARD can send them to queue 1, where they experience negligible waiting time.

In fact, CARD has *optimal* mean response time in heavy traffic, i.e. $C_{\mathrm{CARD}} = C_* := \inf_\pi C_\pi$, answering Question 2.11(b). Roughly speaking, this is because giving all small jobs negligible response time is the best possible outcome. We'll make this more precise when we prove a lower bound on $C_*$ in Section 2.4.

To figure out the details of how CARD should work, we need to better understand closed-loop control. In particular, it is not yet clear how to implement closed-loop control while also dispatching jobs based on their size. As such, we'll spend Section 2.3 studying closed-loop control in a simpler context, namely analyzing LWL, with an eye towards takeaways that will help with designing CARD [Q. 2.17].

## 2.3 Analyzing LWL

The aim of this section is to rigorously analyze LWL while, as discussed at the end of Section 2.2.3, building understanding of closed-loop control in dispatching.

*Question 2.17.* Consider designing closed-loop dispatching policies in the M/G/2/dispatch.

(a) What types of stationary distributions on the system state $(W_1, W_2)$ are achievable with closed-loop control? Given such a distribution, how do we design a dispatching policy that achieves it?

(b) How do we combine closed-loop and size-based dispatching? Can we control the system state with policies that are only "partially" closed-loop?

With these questions in mind, we'll analyze not just LWL, but also an *LWL-$q$* variant that mixes LWL with random dispatching [§ 2.3.4]. We'll see that LWL, and even LWL-$q$, cause the system state distribution to concentrate on states where the work gap $G$ is small, an instance of a more general phenomenon called *state-space collapse* Section 2.3.2.

For instructive purposes, we sometimes assume throughout this section that $S$ is light-tailed, namely that there exists $\eta > 0$ such that such that $\mathbf{E}[e^{\eta S}] < \infty$. We make this assumption because it enables one to prove tighter concentration bounds on the gap $G$, namely an exponential concentration. This is not essential for showing $C_{\mathrm{LWL}} = C_{\mathrm{M/G/1}}$, but it is helpful as a warmup for CARD's analysis in Section 2.5, which involves proving exponential concentration using a similar strategy.

### 2.3.1 Bounding response time in terms of the work gap

**Definition 2.18.** The *idleness process* of a work process $W$, denoted $I_W$ or simply $I$, is

$$I(t) \coloneqq 1 - \mathrm{D}W(t).$$

**Theorem 2.19.** *Let $W$ be a work process that has M/G arrivals, and suppose $\mathbf{E}[S^2] < \infty$ [Asm. 2.6(b)]. Then*

$$\mathbf{E}[W] = \frac{\rho\,\mathbf{E}[S_{\mathrm{e}}] + \mathbf{E}[IW]}{1-\rho} = \mathbf{E}[W_{\mathrm{M/G/1}}] + \frac{\mathbf{E}[IW]}{1-\rho},$$

*where $W_{\mathrm{M/G/1}}$ is an M/G/1 work process with the same M/G arrival parameters [Ntn. 1.17(e)].*

*Proof.* Using RCL 1.22 on $W^2$ yields, along the same lines as in Section 1.3.2,

$$\begin{aligned} 0 &= 2\,\mathbf{E}[W \cdot \mathrm{D}W] + \lambda\,\mathbf{E}_A[(W+S)^2 - W^2] \\ &= -2\,\mathbf{E}[W] + 2\,\mathbf{E}[IW] + 2\lambda\,\mathbf{E}_A[SW] + \lambda\,\mathbf{E}_A[S^2] \\ &= -2\,\mathbf{E}[W] + 2\,\mathbf{E}[IW] + 2\rho\,\mathbf{E}[W] + \lambda\,\mathbf{E}[S^2], \end{aligned}$$

where the last line uses PASTA 1.27, the independence of $S$ from $W$ [Def. 1.15(b)], and our notation convention for expectations involving a freshly sampled job size $S$ [Ntn. 1.3(f)]. The result follows by solving for $\mathbf{E}[W]$ and applying (1.3) and Definition 1.33. □

What does the $\mathbf{E}[IW]$ term from Theorem 2.19 look like in the M/G/2/dispatch? Whenever $I$ is nonzero, one of the queues has no work, in which case $W = G$. This means

$$IW = IG = \frac{\mathbb{1}(W_1 = 0) + \mathbb{1}(W_2 = 0)}{2} \cdot G. \tag{2.10}$$

So understanding the $\mathbf{E}[IW]$ term requires understanding the gap $G$.

In fact, understanding the response time of LWL boils down nearly entirely to understanding $G$. In particular, because we always send jobs to the queue with less work,

$$T^{\mathrm{wait}} = 2W_J = W - G.$$

This means

$$\begin{aligned}\mathbf{E}_A[T^{\text{wait}}] &= \mathbf{E}[W_{\text{M/G/1}}] + \frac{\mathbf{E}[IG]}{1-\rho} - \mathbf{E}[G] \\ &= \mathbf{E}[W_{\text{M/G/1}}] + \frac{\mathbf{Cov}[I,G]}{1-\rho},\end{aligned} \tag{2.11}$$

with the covariance framing [6] following from the fact that $\mathbf{E}[I] = 1-\rho$ [Exr. 2.3].

Exactly analyzing $\mathbf{E}[IG]$ is essentially intractable, as doing so would amount to an exact analysis of the intractable (central-queue) M/G/$k$ [Exr. 2.1]. But the following lemma shows that it suffices to prove what amount to tail bounds for $G$, specifically bounds on $\mathbf{E}[(G-x)^+]$.

**Lemma 2.20.** *In the M/G/2/dispatch using any dispatching policy, for all $x \geq 0$,*

$$\frac{\mathbf{E}[IW]}{1-\rho} = \frac{\mathbf{E}[IG]}{1-\rho} \leq x + \frac{\mathbf{E}[(G-x)^+]}{1-\rho}.$$

*Proof.* We have $IW = IG$ from (2.10). We then compute

$$\begin{aligned}\mathbf{E}[IG] &\leq \mathbf{E}\big[I\big(x + (G-x)^+\big)\big] \\ &= x\,\mathbf{E}[I] + \mathbf{E}[I(G-x)^+] \\ &\leq x\,\mathbf{E}[I] + \mathbf{E}[(G-x)^+]\end{aligned}$$

and use the fact that $\mathbf{E}[I] = 1-\rho$ [Exr. 2.3]. □

**Lemma 2.21.** *In the M/G/2/dispatch using LWL, for all $x \geq 0$,*

$$\left|\mathbf{E}_A[T^{\text{wait}}_{\text{LWL}}] - \mathbf{E}_A[T^{\text{wait}}_{\text{M/G/1}}]\right| \leq x + \frac{\mathbf{E}[(G-x)^+]}{1-\rho},$$

*and so*

$$\left|\mathbf{E}_A[T_{\text{LWL}}] - \mathbf{E}_A[T_{\text{M/G/1}}]\right| \leq x + \frac{\mathbf{E}[(G-x)^+]}{1-\rho} + \mathbf{E}[S].$$

*Proof.* Combining (2.11) and Lemma 2.20 yields

$$\left|\mathbf{E}_A[T^{\text{wait}}] - \mathbf{E}[W_{\text{M/G/1}}]\right| \leq \max\left\{\mathbf{E}[G], x + \frac{\mathbf{E}[(G-x)^+]}{1-\rho}\right\},$$

and we can drop the $\mathbf{E}[G]$ branch because

$$\mathbf{E}[G] \leq x + \mathbf{E}[(G-x)^+] \leq x + \frac{\mathbf{E}[(G-x)^+]}{1-\rho}.$$

The result then follows from (2.1), the following observations about waiting and response time in the M/G/1 under FCFS (where $t \in A$), and PASTA 1.27:

$$T_{\text{M/G/1}}(t) = T^{\text{wait}}_{\text{M/G/1}}(t) + S(t), \qquad T^{\text{wait}}_{\text{M/G/1}}(t) = W_{\text{M/G/1}}(t).$$ □

### 2.3.2 Bounding the work gap

With Lemma 2.21 in hand, proving $\mathbf{E}_A[T_{\mathrm{LWL}}] \approx \mathbf{E}_A[T_{\mathrm{LWL}}]$ amounts to proving bounds on the work gap $G$ under LWL, specifically bounding $\mathbf{E}[(G-x)^+]$. By a Chernoff bound argument [Exr. 2.4], we can bound $\mathbf{E}[(G-x)^+]$ by an exponentially decreasing function of $x$ if we can bound $\mathbf{E}[e^{\eta G}]$ for some $\eta > 0$.

**Lemma 2.22.** *Consider an M/G/2/dispatch using LWL. For all $\theta > 0$ such that $\mathbf{E}[e^{\theta S}] < \infty$,*

$$\mathbf{E}[e^{\theta G}] \le \frac{\mathbf{E}[e^{\theta S} - e^{-\theta S}]}{\mathbf{E}[1 - e^{-\theta S}]} = \frac{\mathbf{E}[e^{\theta S_{\mathrm{e}}}]}{\mathbf{E}[e^{-\theta S_{\mathrm{e}}}]} + 1.$$

*Proof.* The main idea is to use RCL 1.22 on $e^{\theta G}$. We first give a proof that assumes $\mathbf{E}[e^{\theta G}] < \infty$, then show how to adapt the proof to work without this extra assumption. Note that the two expressions given for the bound are equal by Theorem 1.32(b).

As usual, we start by understanding the derivative and jumps of $e^{\theta G}$, which amounts to understanding the derivative and jumps of $G$.

- It turns out that for our purposes, it suffices to observe $\mathrm{D}G \in \{-1/2, 0\}$. This holds because between arrivals, $G$ only changes if the two servers are completing work at different rates. But this only happens when one server is idle and the other is busy, which decreases the gap $G$ at rate $1/2$.
- The gap only jumps when a job arrives, so $\mathrm{J}G \subseteq A$.[2]
- At arrival times $t \in A$, the gap changes from $G(t-) = G(t)$ to $G(t+) = |G(t) - S|$, where $S$, the size of the arriving job [Ntn. 1.17(c)], is independent of $G(t)$.

Using RCL 1.22 on $e^{\theta G}$, the above observations, and PASTA 1.27 yields

$$0 = \theta\, \mathbf{E}[e^{\theta G} \cdot \mathrm{D}G] + \lambda\, \mathbf{E}_A[e^{\theta |G-S|} - e^{\theta G}] \le \lambda\, \mathbf{E}[e^{\theta |G-S|} - e^{\theta G}], \tag{2.12}$$

where we have used the fact that

$$\mathbf{E}[|e^{\theta |G-S|} - e^{\theta G}|] < \max\{\mathbf{E}[e^{\theta S}], \mathbf{E}[e^{\theta G}]\} < \infty.$$

The next step is to isolate $\mathbf{E}[e^{\theta G}]$ on the right-hand side of (2.12). If we had $e^{\theta(G-S)}$ instead of $e^{\theta|G-S|}$, because $S$ is independent of $G$, we would have

$$\mathbf{E}[e^{\theta(G-S)} - e^{\theta G}] = -\mathbf{E}[e^{\theta G}]\, \mathbf{E}[1 - e^{-\theta S}].$$

So let's add and subtract $e^{\theta(G-S)}$ from (2.12), obtaining

$$0 \le \mathbf{E}[e^{\theta(G-S)} - e^{\theta G} + e^{\theta|G-S|} - e^{\theta(G-S)}]. \tag{2.13}$$

[2] In fact, one can show that $\mathrm{J}G = A$ almost surely. The only way for an arrival to cause no jump is if the arrival has size $2G$. But one can show that $(G \mid G \neq 0)$ has a continuous distribution under M/G arrivals, so this happens with probability 0.

After some rearranging, we get

$$\begin{aligned}\mathbf{E}[e^{\theta G}]\,\mathbf{E}[1-e^{-\theta S}] &\le \mathbf{E}[e^{\theta|G-S|}-e^{\theta(G-S)}]\\ &= \mathbf{E}[(e^{\theta(S-G)}-e^{\theta(G-S)})^+]\\ &= \mathbf{E}[\sup_{g\ge 0}(e^{\theta(S-g)}-e^{\theta(g-S)})^+]\\ &= \mathbf{E}[e^{\theta S}-e^{-\theta S}],\end{aligned} \tag{2.14}$$

which yields the desired formula.

How do we adapt the above argument to work without assuming $\mathbf{E}[e^{\theta G}] < \infty$? The idea is, for generic $m > 0$, to use RCL 1.22 on $e^{\theta \min\{G,m\}}$ instead of $e^{\theta G}$, use this to obtain a bound that involves some sort of truncation by $m$, then take the $m \to \infty$ limit.

Using RCL 1.22 on $e^{\theta \min\{G,m\}}$, by the same reasoning as (2.12), we get

$$0 \le \mathbf{E}[e^{\theta \min\{|G-S|,m\}} - e^{\theta \min\{G,m\}}].$$

The key idea is to notice that if $G \ge m$, then $\min\{G,m\}$ can only jump down. And if $G < m$, then simplifications like $\min\{G,m\} = G$ become possible. We compute

$$\begin{aligned}0 &\le \mathbf{E}[e^{\theta \min\{|G-S|,m\}} - e^{\theta \min\{G,m\}}]\\ &\le \mathbf{E}\Big[\big(e^{\theta \min\{|G-S|,m\}} - e^{\theta \min\{G,m\}}\big)\,\mathbb{1}(G<m)\Big]\\ &= \mathbf{E}\Big[\big(-e^{\theta G}(1-e^{-\theta S}) + (e^{\theta\min\{S-G,m\}}-e^{\theta(G-S)})^+\big)\,\mathbb{1}(G<m)\Big]\\ &\le \mathbf{E}\Big[-e^{\theta G}(1-e^{-\theta S})\,\mathbb{1}(G<m) + (e^{\theta(S-G)}-e^{\theta(G-S)})^+\Big]\\ &\le -\mathbf{E}[e^{\theta G}\,\mathbb{1}(G<m)]\,\mathbf{E}[1-e^{-\theta S}] + \mathbf{E}[e^{\theta S}-e^{-\theta S}],\end{aligned}$$

where the last step uses the independence of $S$ and $G$ and the reasoning leading up to (2.14). Rearranging, we get

$$\mathbf{E}[e^{\theta G}\,\mathbb{1}(G<m)] \le \frac{\mathbf{E}[e^{\theta S}-e^{-\theta S}]}{\mathbf{E}[1-e^{-\theta S}]}.$$

The result follows from monotone convergence theorem by taking $m \to \infty$. □

**Corollary 2.23.** *Consider an M/G/2/dispatch using LWL. For all $\theta > 0$ such that $\mathbf{E}[e^{\theta S}] < \infty$ and all $x > 0$,*

$$\mathbf{P}[G > x] \le \frac{\mathbf{E}[e^{\theta S}-e^{-\theta S}]}{\mathbf{E}[1-e^{-\theta S}]}e^{-\theta x}, \qquad \mathbf{E}[(G-x)^+] \le \frac{\mathbf{E}[e^{\theta S}-e^{-\theta S}]}{e\theta\,\mathbf{E}[1-e^{-\theta S}]}e^{-\theta x},$$

*so $\mathbf{P}[G > x] \lesssim e^{-\theta x}$ and $\mathbf{E}[(G-x)^+] \lesssim e^{-\theta x}$ as $x \to \infty$.*

*Proof.* Both bounds follow quickly from Lemma 2.22. The bound on $\mathbf{P}[G > x]$ is the standard Chernoff bound that follows from Markov's inequality, namely

$$\mathbf{P}[G > x] = \mathbf{P}[e^{\theta G} > e^{\theta x}] \le \mathbf{E}[e^{\theta G}]\,e^{-\theta x}.$$

From this, a weaker bound on $\mathbf{E}[(G-x)^+]$ follows from the tail integral formula:

$$\mathbf{E}[(G-x)^+] = \int_x^\infty \mathbf{P}[G > y]\,\mathrm{d}y \le \mathbf{E}[e^{\theta G}]\int_x^\infty e^{-\theta y}\,\mathrm{d}y = \frac{\mathbf{E}[e^{\theta G}]}{\theta}e^{-\theta x}.$$

The desired bound, which is has an extra $1/e$ factor, follows from Exercise 2.4. □

### State-space collapse under LWL

Notice that the bounds in Corollary 2.23 depend on the job size distribution $S$, but *not* on the load $\rho$. That is, although the distribution of the gap $G$ changes as $\rho \longrightarrow 1$, it remains uniformly bounded.

Here's a way to visualize this. We can view $G$ as measuring the (1-norm) distance from the system state $(W_1, W_2)$ to its projection onto the diagonal $\mathbb{D} := \{(x,x) : x \in \mathbb{R}\}$, namely $(W/2, W/2)$. So we can interpret Corollary 2.23 as saying that the distance between $(W_1, W_2)$ and $\mathbb{D}$ remains stochastically bounded as $\rho \longrightarrow 1$, even though other aspects of the system's state, namely $W$, grow unboundedly.

In light of this, we say that LWL induces *state-space collapse* to the diagonal $\mathbb{D}$ in heavy traffic. In general, state-space collapse refers to when a system state's distance to some subset of states remains stochastically bounded in some limiting regime, usually while other distances grow unboundedly in the same limit. Some sources reserve the term for *exponential* tail bounds on the distance, which is the case in Corollary 2.23. But we use the term more broadly. For instance, when $S$ is heavy-tailed, we still get a stochastic bound on the gap $G$ under LWL, though its tail can be far from exponential [Exr. 2.10].

## 2.3.3 Bounding LWL's mean response time

**Theorem 2.24.** *Consider an M/G/2/dispatch using LWL, and suppose $\mathbf{E}[e^{\theta S}] < \infty$ for some $\theta > 0$. Then*

$$\left|\mathbf{E}_A[T_{\mathrm{LWL}}] - \mathbf{E}_A[T_{\mathrm{M/G/1}}]\right| \le \frac{1}{\theta}\left(\log\frac{1}{1-\rho} + \log\frac{\mathbf{E}[e^{\theta S} - e^{-\theta S}]}{\mathbf{E}[1 - e^{-\theta S}]}\right) + \mathbf{E}[S].$$

*This means as $\rho \longrightarrow 1$,*

$$\left|\mathbf{E}_A[T_{\mathrm{LWL}}] - \mathbf{E}_A[T_{\mathrm{M/G/1}}]\right| \lesssim \log\frac{1}{1-\rho},$$

*and thus $\mathbf{E}_A[T_{\mathrm{LWL}}] \approx \mathbf{E}_A[T_{\mathrm{M/G/1}}]$ and $C_{\mathrm{LWL}} = C_{\mathrm{M/G/1}} = \mathbf{E}[S_{\mathrm{e}}]$.*

*Proof.* Plugging Corollary 2.23 into Lemma 2.21 yields

$$\left|\mathbf{E}_A[T_{\mathrm{LWL}}] - \mathbf{E}_A[T_{\mathrm{M/G/1}}]\right| \le x + \frac{m}{e\theta(1-\rho)}e^{-\theta x} + \mathbf{E}[S],$$

where $m := \mathbf{E}[e^{\theta S} - e^{-\theta S}]/\mathbf{E}[1 - e^{-\theta S}]$. This is minimized when

$$x = \frac{1}{\theta}\left(\log \frac{m}{1-\rho} - 1\right),$$

yielding

$$\left|\mathbf{E}_A[T_{\mathrm{LWL}}] - \mathbf{E}_A[T_{\mathrm{M/G/1}}]\right| \le \frac{1}{\theta}\log\frac{m}{1-\rho} + \mathbf{E}[S]. \qquad \square$$

### 2.3.4 Extending the analysis to partially randomized LWL-$q$

**Policy 2.25.** *Random* is the dispatching policy that dispatches each job uniformly at random. That is, $J_{\mathrm{Random}}$ is the i.i.d. mark process on $A$ where for all $t \in A$,

$$J_{\mathrm{Random}}(t) \sim \mathrm{Unif}\{1, \dots, k\}.$$

**Policy 2.26.** *LWL-$q$* is the dispatching policy that for each arrival follows LWL with probability $q$, following Random otherwise. That is, at arrival times $t \in A$,

$$J_{\mathrm{LWL}\text{-}q}(t) = \begin{cases} J_{\mathrm{LWL}}(t) & \text{with probability } q \\ J_{\mathrm{Random}}(t) & \text{with probability } 1-q. \end{cases}$$

*Notation 2.27.* We have introduced yet another mild abuse of notation in Policy 2.26.

- In Definition 2.3(a) and Notation 2.5, we write $J$ or $J_\pi$ for the actual dispatch index in an M/G/$k$/dispatch under a generic policy $\pi$.
- But in Policy 2.26, we write $J_{\mathrm{LWL}}$ and $J_{\mathrm{Random}}$ for what the dispatch index *would be* under LWL and Random, respectively, even though the actual policy in use is LWL-$q$.

This notation should be understood in the obvious way. For example, given some other dispatching policy, which induces some queue work processes $W_1, \dots, W_k$, we let $J_{\mathrm{LWL}}(t) := \arg\min_i W_i(t)$, even if some policy other than LWL is in use.

An alternative would be to give some other name to $t \mapsto \arg\min_i W_i(t)$, then define LWL to be the policy under which $J$ is that process. Instead, we simply ensure it is clear from context what policy is actually in use whenever notation like $J_{\mathrm{LWL}}$ appears.

Under LWL-$q$, by PASTA 1.27 and (2.1) we have

$$\mathbf{E}_A[T_{\mathrm{LWL}\text{-}q}] = \mathbf{E}[W] - q\,\mathbf{E}[G] + 2\,\mathbf{E}[S].$$

This is because with probability $q$, we follow LWL, contributing $-G$ to waiting time; and otherwise, we follow Random, contributing 0 to waiting time in expectation. By the same reasoning as Lemma 2.21, we obtain

$$\left|\mathbf{E}_A[T_{\mathrm{LWL}\text{-}q}] - \mathbf{E}_A[T_{\mathrm{M/G/1}}]\right| \le x + \frac{\mathbf{E}[(G-x)^+]}{1-\rho} + \mathbf{E}[S]. \tag{2.15}$$

So understanding LWL-$q$ boils down to understanding $G$. We'll soon see that the bounds we get on $G$ get larger the smaller $q$ is. Our is to characterize how small $q$ can be as a function of the load $\rho$ in the heavy traffic limit while still maintaining $\mathbf{E}_A[T_{\mathrm{LWL}\text{-}q}] \approx \mathbf{E}_A[T_{\mathrm{M/G/1}}]$.

*Remark 2.28.* Equation (2.15) leaves open the possibility that LWL-$q$'s mean response time could benefit from having a larger gap. However, Exercise 2.14 implies $\mathbf{E}_A[T_\pi] \gtrapprox \mathbf{E}_A[T_{\mathrm{M/G/1}}]$ for a class of policies $\pi$ that includes LWL-$q$.

We can also see $\mathbf{E}_A[T_{\text{LWL-}q}] \gtrapprox \mathbf{E}_A[T_{\mathrm{M/G/1}}]$ directly. The lower bound

$$\mathbf{E}_A[T_{\text{LWL-}q}] = \mathbf{E}[W] - q\,\mathbf{E}[G] + 2\,\mathbf{E}[S] \geq \mathbf{E}[W] - q\,\mathbf{E}[W] \geq (1-q)\,\mathbf{E}[W_{\mathrm{M/G/1}}]$$

shows that we don't benefit from a large gap if $q$ is small, and in particular not if we take $q \to 0$ as $\rho \to 1$. And we'll soon see that if $q$ remains constant as $\rho \to 1$, then $\mathbf{E}[G]$ remains bounded by a constant. Either way, we get $\mathbf{E}_A[T_{\text{LWL-}q}] \gtrapprox \mathbf{E}[W_{\mathrm{M/G/1}}] \approx \mathbf{E}_A[T_{\mathrm{M/G/1}}]$.

### Bounding LWL-$q$'s work gap and mean response time

*Notation 2.29.* We say that a statement holds for $x$ *sufficiently close* to $y$ if there exists an interval of nonzero length $Z \ni y$ such that the statement holds for all $x \in Z \setminus \{y\}$. Whether $Z$ should contain points below, above, or on both sides of $y$ will be clear from context. *Sufficiently small* means sufficiently close to 0, and *sufficiently large* means sufficiently close to $\infty$.

**Lemma 2.30.** *Consider an M/G/2/dispatch using LWL-q. Suppose there exists $\eta > 0$ such that $\mathbf{E}[e^{\eta S}] < \infty$, and let*

$$r(\theta) := \frac{\mathbf{E}[e^{\theta S} + e^{-\theta S} - 2]}{\mathbf{E}[e^{\theta S} - e^{-\theta S}]} = \frac{\mathbf{E}[e^{\theta S_{\mathrm{e}}} - e^{-\theta S_{\mathrm{e}}}]}{\mathbf{E}[e^{\theta S_{\mathrm{e}}} + e^{-\theta S_{\mathrm{e}}}]}.$$

(a) *For all $\theta \in [0, \eta]$ such that $r(\theta) < q$,*

$$\mathbf{E}[e^{\theta G}] \leq \frac{1+q}{q - r(\theta)}.$$

(b) *For all $\theta \in [0, \eta]$,*

$$r(\theta) \leq \mathbf{E}[e^{\theta S_{\mathrm{e}}} - 1].$$

*In particular, for sufficiently small $\theta$,*

$$r(\theta) \leq 2\theta\,\mathbf{E}[S_{\mathrm{e}}].$$

(c) *For sufficiently small $q$ and all $x \geq 0$,*

$$\mathbf{E}\left[\exp\left(\frac{q}{4\,\mathbf{E}[S_{\mathrm{e}}]}G\right)\right] \leq \frac{4}{q}, \qquad \mathbf{E}[(G-x)^+] \leq \frac{16\,\mathbf{E}[S_{\mathrm{e}}]}{eq^2}\exp\left(-\frac{q}{4\,\mathbf{E}[S_{\mathrm{e}}]}x\right).$$

*Specifically, there exists $q' > 0$, which depends only on $S$, such that the bounds hold for $q \in (0, q']$; and for $q \in (q', 1]$, the bounds hold with $q$ replaced by $q'$.*

*Proof.*

(a) The approach is very similar to the proof of Lemma 2.22, so we explain just the key differences. Applying RCL 1.22 to $e^{\theta G}$ and following the reasoning from the proof of Lemma 2.22

$$\begin{aligned} 0 &= \theta\,\mathbf{E}[e^{\theta G}\cdot \mathrm{D}G] + \lambda\,\mathbf{E}_A\Big[e^{G+S}\,\mathbb{1}(J = \arg\max_i W_i) + e^{\theta|G-S|}\,\mathbb{1}(J = \arg\min_i W_i) - e^{\theta G}\Big] \\ &\le \lambda\,\mathbf{E}\left[\frac{1-q}{2}e^{\theta(G+S)} + \frac{1+q}{2}e^{\theta|G-S|} - e^{\theta G}\right], \end{aligned}$$

where the second step uses the fact that the event $J = \arg\min_i W_i$ is independent of $G$ and $S$. Bounding

$$e^{\theta|G-S|} = e^{\theta(G-S)} + (e^{\theta(S-G)} - e^{\theta(G-S)})^+ \le e^{\theta(G-S)} + e^{\theta S} - e^{-\theta S}$$

and using the independence of $G$ and $S$, this becomes

$$\mathbf{E}[e^{\theta G}]\,\mathbf{E}[(1-q)e^{\theta S} - (1+q)e^{-\theta S} - 2] \le \mathbf{E}[e^{\theta S} - e^{-\theta S}].$$

Provided $\mathbf{E}[(1-q)e^{\theta S} - (1+q)e^{-\theta S} - 2] > 0$, or equivalently $r(\theta) < q$, this rearranges to the desired bound in terms of $r(\theta)$. The two expressions given for $r(\theta)$ are equal by Theorem 1.32(b).

Above, we implicitly assumed that $\mathbf{E}[e^{\theta G}] < \infty$ when applying RCL 1.22 to $e^{\theta G}$. But, as in the proof of Lemma 2.22, essentially the same computation with $G$ replaced by $\min\{G, m\}$ gives a proof that doesn't rely on this assumption.

(b) The first bound on $r(\theta)$ follows by Jensen's inequality, specifically $\mathbf{E}[e^{-\theta S_e}] \ge 1/\mathbf{E}[e^{\theta S_e}]$, and the fact that $\frac{x-1/x}{x+1/x} \le x - 1$ for all $x \ge 1$. The second bound follows by convexity of $\theta \mapsto \mathbf{E}[e^{\theta S_e}]$.

(c) The first bound follows by plugging in $\theta = \frac{q}{4\,\mathbf{E}[S_e]}$, and the second follows from the first by Exercise 2.4. The last statement about substituting in sufficiently small $q'$ if $q$ isn't small enough holds because the bound in (a) is monotonic in $q$. □

**Theorem 2.31.** *Consider an M/G/2/dispatch using LWL-q, and suppose* $\mathbf{E}[e^{\theta S}] < \infty$ *for some* $\theta > 0$.

(a) *For sufficiently small q,*

$$\Big|\mathbf{E}_A[T_{\text{LWL-}q}] - \mathbf{E}_A[T_{\text{M/G/1}}]\Big| \le \frac{4}{q}\left(\log\frac{1}{1-\rho} + \log\frac{4}{q}\right).$$

(b) *If q remains fixed as* $\rho \longrightarrow 1$, *or more generally if* $q \sim 1$, *then*

$$\Big|\mathbf{E}_A[T_{\text{LWL-}q}] - \mathbf{E}_A[T_{\text{M/G/1}}]\Big| \lesssim \log\frac{1}{1-\rho}.$$

*and thus* $\mathbf{E}_A[T_{\text{LWL-}q}] \approx \mathbf{E}_A[T_{\text{M/G/1}}]$ *and* $C_{\text{LWL-}q} = C_{\text{M/G/1}} = \mathbf{E}[S_e]$.

(c) *If we have, as $\rho \longrightarrow 1$,*

$$q \gg (1-\rho)\log\frac{1}{1-\rho},$$

*then* $\mathbf{E}_A[T_{\text{LWL-}q}] \approx \mathbf{E}_A[T_{\text{M/G/1}}]$ *and* $C_{\text{LWL-}q} = C_{\text{M/G/1}} = \mathbf{E}[S_\text{e}]$.

*Proof.* Combining (2.15) and Lemma 2.30(c) gives a bound that is minimized when

$$x = \frac{4\,\mathbf{E}[S_\text{e}]}{q}\left(\log\frac{4}{q(1-\rho)} - 1\right),$$

yielding (a). This immediately implies (b), and (c) follows because $\mathbf{E}_A[T_{\text{M/G/1}}] \sim \frac{1}{1-\rho}$. □

## 2.4 Lower bound on optimal dispatching

*Coming soon!*

## 2.5 Designing a near-optimal dispatching policy

*Coming soon!*

**Policy 2.32.** *Controlled Asymmetry Reduces Delay (CARD)* is a pretty cool dispatching policy. *More coming soon!*

## 2.6 Exercises

### 2.6.1 Basic properties of LWL

**Exercise 2.1.** Prove that for an *arbitrary* arrival sequence of jobs, using LWL dispatching in a $k$-server system is equivalent to using a first-come first-served central queue. *Hint:* When a job arrives to a central-server system, can you tell from the system state which server will eventually serve it?

**Exercise 2.2.** Consider a GD/G/2/dispatch using LWL, and recall the gap is $G = |W_1 - W_2|$ [Def. 2.3(d)]. Let $\hat{G}$ be the process on $[0, \infty)$ that satisfies

$$\hat{G}(0) = G(0), \tag{2.16}$$

$$\mathrm{D}\hat{G}(t) = 0, \tag{2.17}$$

$$\hat{G}(t+) = \max\left\{|\hat{G}(t-) - S(t)|, S(t)\right\} \quad \text{for } t \in A, \tag{2.18}$$

where $S = (t \mapsto S(t))$ is the arrival process and $A$ is its set of arrival times [Def. 1.36].

(a) Show $\hat{G}(t) \geq |G(t)|$ for all $t \geq 0$.

(b) What can you conclude about $G$ when the size distribution is bounded?

(c) *Challenge!* Define a version of $\hat{G}$ that is jointly stationary with $S$. Your definition should still satisfy (2.17) and (2.18), but it need not satisfy (2.16). *Hint:* It's not right to start with $\hat{G}(0) = 0$, because $G(0) > 0$ almost surely. But you would still eventually get to some time $t$ when $\hat{G}(t) \geq G(t)$. What if instead of $\hat{G}(0) = 0$, you started with $\hat{G}(-100) = 0$? It's okay if your definition only works almost surely.

The motivation for this problem is that one thing that makes $G$ so difficult to analyze exactly is that it is affected by when the servers are idle, which is intractable to capture exactly. The idea is that $\hat{G}$ is an upper bound on $G$ that is not affected by idleness.

### 2.6.2 Filling in details

**Exercise 2.3.** Let $I$ be the idleness process of a work process that has M/G arrivals. Show

$$\mathbf{E}[I] = 1 - \rho.$$

**Exercise 2.4.** Let $V$ be a random variable. Show that for any $\theta > 0$,

$$\mathbf{E}[(V - x)^+] \leq \frac{\mathbf{E}[e^{\theta V}]}{e\theta} e^{-\theta x}.$$

*Hint:* The usual Chernoff bound uses the fact that $\mathbb{1}(V > x) \leq e^{V-x}$. What similar fact can you use here?

**Exercise 2.5.** Complete the computation of $\mathbf{E}_A[T_{\text{SITA-E}}]$, obtaining (2.6) and (2.7).

**Exercise 2.6.** Consider an M/G/2/dispatch using SITA-O with a continuous size distribution $S$ [Asm. 2.6(c)]. Let $r$, which depends on the load $\rho$, be the optimizing threshold, and recall $q$ is the load median [Def. 2.14(b)]. We must have $r \to q$ as $\rho \to 1$, or else we would make one of the queues unstable [Asm. 1.18].

(a) Show

$$\begin{aligned} \mathbf{P}[S < r] &\approx \mathbf{P}[S < q], \\ \mathbf{E}[S\,\mathbb{1}(S < r)] &\approx \mathbf{E}[S\,\mathbb{1}(S < q)], \\ \mathbf{E}[S^2\,\mathbb{1}(S < r)] &\approx \mathbf{E}[S^2\,\mathbb{1}(S < q)] \end{aligned}$$

(b) Use (a) to express $C_{\text{SITA-O}}$ as the solution to a single-variable optimization problem. *Hint:* It might seem like (a) implies there are no choices to be made. However, there is one part of the mean response time formula where simply substituting a limiting value from (a) is not valid! This is where the decision variable comes from.

(c) Compute $C_{\text{SITA-O}}$.

### 2.6.3 Dispatching as altering the arrival process

**Exercise 2.7.** The next few exercises ask you to think about dispatching systems from the perspective of a single queue. From one queue's perspective, being in a dispatching system with M/G arrivals looks like being a single-server queue with a more complicated arrival process, such as G/G arrivals. This means the Kingman bound (1.12) can give us insight into dispatching systems. To that end, in this problem, you'll show the Kingman bound for the G/G/1 is actually tight in heavy traffic, namely (2.19) below. That said, free to skip this problem and take (2.19) as given for now.

Consider the heavy-traffic G/G/1 with $\mathbf{E}[R^2] < \infty$ and $\mathbf{E}[S^2] < \infty$. More precisely, consider the $\rho \to 1$ limit with fixed size distribution $S$ and linearly scaled interarrival time distribution $R$, meaning there exists a fixed distribution $R'$ such that $R := R'/\rho$ and $\mathbf{E}[R'] = \mathbf{E}[S]$. Your task is to show

$$\mathbf{E}_A[W] \approx \frac{\mathbf{Var}[R - S]}{2\,\mathbf{E}[R - S]} \approx \frac{\frac{1}{2}(c_R^2 + c_S^2)}{1 - \rho}\,\mathbf{E}[S]. \tag{2.19}$$

(a) Show $\mathbf{E}_A\Big[\big((R - S - W)^+\big)^2\Big]$, the intractable term from the G/G/1 $\mathbf{E}_A[W]$ formula (1.9), can be expressed as

$$\mathbf{E}_A\Big[\big((R - S - W)^+\big)^2\Big] = 2\,\mathbf{E}[IA_1],$$

where $I = \mathbb{1}(W = 0)$ is the idleness process of the G/G/1 work process $W$. *Hint:* Use PIF 1.25, and recall what variables have distributions $R$ and $S$ under $\mathbf{P}_A[\cdot]$.

(b) Show (2.19) by showing $\lim_{\rho\to 1} \mathbf{E}[IA_1] = 0$. If you like, you may assume an extra condition on $R$, but the result can be shown without such a condition. *Hint:* For example, if you assume $\mathbf{E}[R^3] < \infty$, then there is a solution using Cauchy-Schwarz. More generally, use the fact that $R := R'/\rho$ is uniformly integrable as a function of $\rho$.

(c) Show $\mathbf{E}[W] \approx \mathbf{E}_A[W]$ in heavy traffic.

**Exercise 2.8.** Consider an M/G/$k$/dispatch using *Round Robin (RR)*, which dispatches jobs by cycling through servers in a fixed order: $1, 2, \ldots, k, 1, 2, \ldots$. Formally,

$$J_{i+1} \equiv J_i + 1 \mod k.$$

Give upper bounds on the waiting time $k\,\mathbf{E}_A[W_J]$ for the following systems.

(a) The M/G/2/dispatch using RR. *Hint:* This is secretly a G/G/1 question.

(b) The M/G/$k$/dispatch using RR. *Hint:* This is still secretly a G/G/1 question.

(c) *Challenge!* The M/G/$k$/dispatch using a variant of RR where within each "cycle" of sending one job to each server, the order of the servers is freshly sampled uniformly at random. *Hint:* This is secretly a *GD/G/1* question, not simply a G/G/1 question! It's related to Exercise 1.22.

Throughout, consider how your answers relate to the "work ≈ intensity × variability" intuition from Section 1.4.

**Exercise 2.9.** Consider an M/G/$k$/dispatch using LWL. Each queue $i$ work process $W_i$ is a speed $1/k$ standard work process. In heavy traffic, by symmetry and (2.3),

$$\mathbf{E}[W_i] \approx \frac{1}{k}\,\mathbf{E}[W_{\mathrm{LWL}}] \approx \frac{1}{k}\,\mathbf{E}[W_{\mathrm{M/G/1}}].$$

Compare this to the Kingman bound on $\mathbf{E}[W_{\mathrm{G/G/1}}]$ in (1.12), using a G/G/1 with the same size distribution $S$ and load $\rho$. Given the fact that the Kingman bound is tight in heavy traffic [Exr. 2.7], what would $c_R^2$ need to be to have $\mathbf{E}[W_{\mathrm{G/G/1}}] \approx \mathbf{E}[W_i]$? *Hint:* The hypothetical word “would” is key.

### 2.6.4 Generalizations

**Exercise 2.10.** Consider an M/G/2/dispatch, and suppose $\mathbf{E}[S^2] < \infty$. Do *not* assume other moment or transform bounds on $S$. (Well, aside from $S > 0$, hence $\mathbf{E}[S] > 0$.)

(a) Show $\mathbf{E}[G] \le \mathbf{E}[S_{\mathrm{e}}]$.

(b) Show

$$\mathbf{P}[G > x + y] \le \frac{\mathbf{E}[(S - y)^+]}{\mathbf{E}[\min\{S, x\}]}.$$

*Hint:* Apply RCL 1.22 to $(G - y)^+$. At some point, you'll have a term that you wish was $S$, but it might be smaller. Under what conditions is the term at least $\min\{S, x\}$?

(c) Using your answer to (b), show

$$\mathbf{E}[(G - x)^+] \le \frac{\mathbf{E}[((S - x)^+)^2]}{\mathbf{E}[\min\{S, x\}]}.$$

*Hint:* Integrals and expectations were made to be swapped. Thanks, Tonelli!

(d) Show $\mathbf{E}_A[T_{\mathrm{LWL}}] \approx \mathbf{E}_A[T_{\mathrm{M/G/1}}]$ as $\rho \to 1$. *Hint:* Given that $\mathbf{E}[S^2] < \infty$, how must $\mathbf{E}[(G - x)^+]$ behave as $x \to \infty$?

**Exercise 2.11.** Consider an M/G/$k$/dispatch for general $k \ge 2$.

(a) Suppose $S$ is bounded, meaning there is a maximum size $m$ such that $\mathbf{P}[S \le m] = 1$. Find *universal* constants $a, b > 0$, meaning not depending on $\rho$ or $S$, such that

$$\left|\mathbf{E}_A[T_{\mathrm{LWL}}] - \mathbf{E}_A[T_{\mathrm{M/G/1}}]\right| \le (k - 1)(a\,\mathbf{E}[S] + bm).$$

(b) *Challenge!* Suppose $\mathbf{E}[e^{\theta S}] < \infty$ for some $\theta > 0$. Show $\mathbf{E}_A[T_{\mathrm{LWL}}] \approx \mathbf{E}_A[T_{\mathrm{M/G/1}}]$ as $\rho \to 1$. *Hint:* The approach from Section 2.3 still works, but you'll need to think about more than one gap process. Start by thinking about $G_{ij} = |W_i - W_j|$. Don't worry about proving the tightest possible bounds on terms that are $\lesssim \frac{1}{1-\rho}$.

**Exercise 2.12.** Consider dispatching in the heavy-traffic M/G/$k$/dispatch. Try to answer the following questions *without* doing a formal analysis. Instead, just give heuristic arguments.

(a) What do you think $C_* := \inf_\pi C_\pi$ is? *Hint:* What do you think the ideal system state $(W_1, \ldots, W_k)$ is? What's the largest the the average size of jobs in the system can?

(b) How might you adapt CARD to get a policy that achieves $C_{\text{CARD}} = C_*$? What size classes do you need, and how do you dispatch each size class? How do you think you should set the size thresholds as functions of $k$ and $\rho$? *Hint:* There are ways to generalize CARD to the M/G/$k$/dispatch that include about $2k$ size classes. But there are also simpler ways that use just three classes, though the thresholds depend on $k$.

**Exercise 2.13.** *Open-ended....* Consider dispatching in the heavy-traffic G/G/2/dispatch, G/G/$k$/dispatch, or GD/G/$k$/dispatch. (See Exercise 2.7 for a precise definition of "heavy traffic" for G/G arrivals.)

(a) Can you prove anything about $C_{\text{SITA-E}}$?

(b) Can you prove anything about $C_{\text{LWL}}$?

(c) *Open problem!?* Can you prove anything about $C_{\text{CARD}}$ and $C_*$? Can you find conditions such that $C_{\text{CARD}} = C_*$?

Throughout, it's okay to make additional assumptions about the size and interarrival time distributions. You may use the fact that the Kingman bound (1.12) is tight in heavy traffic [Exr. 2.7].

### 2.6.5 More closed-loop dispatching policies

**Exercise 2.14.** Consider an M/G/2/dispatch using an arbitrary *size-oblivious* dispatching policy, meaning one where the dispatch index $J$ is (conditionally) independent of the arrival's size $S$ (given the system state $(W_1, W_2)$). You will prove that under any size-oblivious policy, we have $\mathbf{E}_A[T] \gtrapprox \mathbf{E}_A[T_{\text{M/G/1}}]$ in heavy traffic.

For arrival times $t \in A$, let

$$B(t) := \begin{cases} -1 & \text{if } J(t) = \arg\min_i W_i \\ +1 & \text{if } J(t) = \arg\max_i W_i, \end{cases}$$

so under $\mathbf{P}_A[\cdot]$, the work gap jumps from $G_{(-)} = G$ to $G_{(+)} = |G + BS|$.

(a) Write an expression for $\mathbf{E}_A[T]$ in terms of $W$, $S$, and $BG$.

(b) Give a load-independent lower bound on $\mathbf{E}_A[BG]$. *Hint:* Try applying RCL 1.22 to some of the usual suspects, e.g. $G$, $G^2$, and $e^{\theta G}$. You'll need to use the fact that $S$ is independent of $B$ and $G$ at arrival times.

(c) Show $\mathbf{E}_A[T] \gtrapprox \mathbf{E}_A[T_{\text{M/G/1}}]$ by giving a lower bound on $\mathbf{E}_A[T]$. One valid bound is

$$\mathbf{E}_A[T] \geq \mathbf{E}_A[T_{\text{M/G/1}}] + \mathbf{E}[S] - \mathbf{E}[S_{\text{e}}],$$

but it's okay if you show a different bound, as long as it is $\approx \mathbf{E}_A[T_{\text{M/G/1}}]$.

**Exercise 2.15.** Consider an M/G/2/dispatch using a dispatching policy $\pi$ with the following property: whenever a job arrives, the (conditional) probability it is sent to the server with less work is at least $1/2$ (given the system state $(W_1, W_2)$ and the arrival's size $S$). That is, we (almost surely) have

$$\mathbf{P}_A\left[J_\pi = \arg\min_i W_i \;\middle|\; W_1, W_2, S\right] \geq \frac{1}{2}. \tag{2.20}$$

You will show that this leads to at most as much work as under the Random policy [Pol. 2.25], for which (2.20) holds with equality. Below, quantities refer to policy $\pi$ unless otherwise stated.

(a) Explain why $W_{\text{Random}}$ has the same distribution as $(W_{\text{M/G/1}})_1 + (W_{\text{M/G/1}})_2$, a sum of two i.i.d. variables with the same distribution as $W_{\text{M/G/1}}$. *Hint:* One strategy is to think about the splitting property of Poisson processes.

(b) Show $\mathbf{E}[W] = \mathbf{E}[W_1] + \mathbf{E}[W_2] \leq \mathbf{E}[W_{\text{Random}}] = 2\,\mathbf{E}[W_{\text{M/G/1}}]$. *Hint:* Try using RCL 1.22 on $W_1^2 + W_2^2$ instead of $W^2 = (W_1 + W_2)^2$. When you encounter difficult terms to do with dispatching decisions, bound them using (2.20). You should get a bound for $\pi$ and an exact result for Random.

(c) Show $\mathbf{E}[W_1^2] + \mathbf{E}[W_2^2] \leq 2\,\mathbf{E}[W_{\text{M/G/1}}^2]$. *Hint:* What should you try using RCL 1.22 on?

(d) Show $\mathbf{E}[W^2] \lesssim \frac{1}{(1-\rho)^2}$. *Hint:* This boils down to showing $\mathbf{E}[W_{\text{M/G/1}}^2] \lesssim \frac{1}{(1-\rho)^2}$. You can use RCL 1.22 on some function of $W_{\text{M/G/1}}$, or you can look at the second derivative of the transform $\mathbf{E}[e^{\theta W_{\text{M/G/1}}}]$ [Exr. 1.12].

**Exercise 2.16.** *Join Below Threshold z (JBT-z)* is the dispatching policy that dispatches as follows whenever a job arrives.

- If any queues have less than $z$ work, send the arrival to a uniformly random queue with less than $z$ work.
- Otherwise, send the arrival to a uniformly random queue.

In this problem, you'll investigate the question: how does $z$ need to scale as a function of load $\rho$ for JBT-$z$ to achieve $\mathbf{E}_A[T_{\text{JBT-}z}] \approx \mathbf{E}_A[T_{\text{M/G/1}}]$ as $\rho \to 1$? Thanks to Exercise 2.14, we can focus just on determining when $\mathbf{E}_A[T_{\text{JBT-}z}] \lessapprox \mathbf{E}_A[T_{\text{M/G/1}}]$.

Throughout, we consider the M/G/2/dispatch at sufficiently high load. You may assume a constant lower bound on $\rho$, e.g. $\rho \geq 2/3$. We also assume the size distribution is light-tailed, specifically $\mathbf{E}[e^{\theta S}] < \infty$ for some $\theta > 0$.

When analyzing LWL, the key was understanding and bounding the work gap $G$. This is because $G$ is the quantity that LWL keeps small by closed-loop control. However, under JBT-$z$, the work gap could be very large. What does JBT-$z$ control?

We can think of JBT-$z$ as trying to keep both queues' work amounts above $z$. This is because when $W_1 < z \leq W_2$, JBT-$z$ sends all arrivals to queue 1, increasing $W_1$ at rate $\rho$. This is faster than the rate at which decreases $W_1$, namely $1/2$ (or 0 if $W_1 = 0$).

It thus seems like JBT-$z$ might keep the quantity $(z - \min_i W_i)^+$ small. However, JBT-$z$ can't generally prevent both queues from occasionally going below $z$ at the same time.

Specifically, when $W_1, W_2 < z$, JBT-$z$ splits arrivals evenly between the queues, so neither queue's work process increases on average (except when at zero).

It turns out that the key quantity that JBT-$z$ controls is

$$V := \min\Big\{\big(z - \min_i W_i\big)^+, (G - z)^+\Big\}.$$

Put another way, $V$ is the taxicab distance between the system state $(W_1, W_2)$ and the region

$$\mathcal{R} := \big\{(x, y) \in [0, \infty)^2 : \min\{x, y\} \geq z \text{ or } |x - y| \leq z\big\},$$

which is illustrated in Figure 2.3. (There are other possible definitions that could work for $V$ and $\mathcal{R}$, but these ones make the computations simpler.)

How does understanding $V$ help us bound mean response time? As usual, bounding mean response time boils down to bounding mean work $\mathbf{E}[W]$, which in turn boils down to bounding $\mathbf{E}[IW]$. And $I$ is related to the $V$, because we can see from Figure 2.3 that whenever there's enough work in the system, a server can be idle only if the state is at least distance $z$ from $\mathcal{R}$, i.e. only if $V = z$. So we'll want to bound $\mathbf{P}[V = z]$, for which bounding $\mathbf{E}[e^{\theta V}]$ and using a Chernoff bound suffices.

Analyze JBT-$z$ in the M/G/2/dispatch by following the steps below. You may assume all the expectations related to $V$ and $W$ you encounter are finite. (As usual, a truncation argument can be used to remove this assumption [Rmk. 1.24, Lem. 2.22].)

(a) Show

$$\mathbf{E}_A[T_{\text{JBT-}z}] \leq \mathbf{E}[W] + 2z + 2\,\mathbf{E}[S].$$

*Hint:* This is a loose bound, so don't overthink it. What would the response time be for a job that was dispatched randomly? What would the response time be for a job that was dispatched to a queue with less than $z$ work?

(b) Give an upper bound on $\mathbf{E}[W]$ in terms of $\mathbf{E}[W_{\text{M/G/1}}]$, $z$, and $\mathbf{E}[W\,\mathbb{1}(V = z)]$. *Hint:* Use a bound like $W \leq cz + (W - cz)^+$ for some constant $c$. Use Figure 2.3 to figure out a good value for $c$.

(c) Give an upper bound on $\mathbf{E}[W\,\mathbb{1}(V = z)]$ in terms of $\mathbf{E}[We^{\theta V}\,\mathbb{1}(V > 0)]$.

(d) Show that for all $\theta$ such that the right-hand side has finite numerator and positive denominator,

$$\mathbf{E}[e^{\theta V}\,\mathbb{1}(V > 0)] \leq \frac{\mathbf{E}[e^{\theta S_e}]}{\mathbf{E}[e^{-\theta S_e}] - \frac{1}{2\rho}}.$$

You may assume that there exists a constant value of $\theta$, which depends only on $S$, such that the above holds for all $\rho \geq 2/3$. *Hint:* Apply RCL 1.22 to $e^{\theta V}$… but you knew that already. Notice that the ways that $V$ can jump differ depending on whether $V = 0$ or $V > 0$. To get a feel for this, draw pictures of possible ways the system state can jump relative to $\mathcal{R}$ [Fig. 2.3], which is likely easier than writing out the formula for $V$. In order to isolate $e^{\theta V}$, you'll need to use tricks similar to how (2.13) adds and subtracts $e^{\theta(G-S)}$. Finally, remember $\lambda\,\mathbf{E}[e^{\theta S} - 1] = \rho\theta\,\mathbf{E}[e^{\theta S_e}]$ [Thm. 1.32(b)].

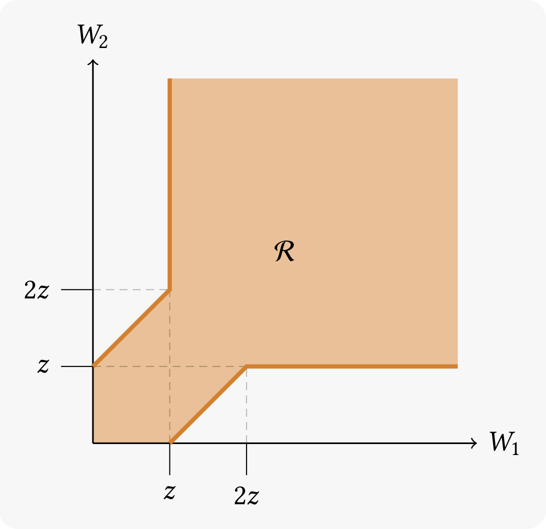


**Figure 2.3.** Region $\mathcal{R}$ of the state space that JBT-$z$ pushes the system state towards. The process $V = \min\{(z - \min_i W_i)^+, (G - z)^+\}$ measures the taxicab distance from the system state $(W_1, W_2)$ to the region $\mathcal{R}$.

(e) Verify that JBT-$z$ satisfies the precondition of Exercise 2.15, which is (roughly) that every job has at least probability $1/2$ of being sent to the queue with less work.

(f) Using Exercise 2.15(d) and the Cauchy-Schwarz inequality, show that there exists a constant $c$, which depends only on $S$, such that if we set $z$ satisfying

$$\frac{c}{\theta} \log \frac{1}{1-\rho} \leq z \ll \frac{1}{1-\rho}$$

as $\rho \longrightarrow 1$, then $\mathbf{E}_A[T_{\text{JBT-}z}] \approx \mathbf{E}[W] \approx \mathbf{E}[W_{\text{M/G/1}}]$.

# Chapter A
# Selected solutions

## A.1 Solutions to exercises from Chapter 1

### Exercise 1.1.

(a) Show that every *rational* $u > 0$ yields the same value in Definition 1.20. If you prefer, you can restrict to *dyadic* rationals, i.e. $u = n/2^k$ for $n, k \in \mathbb{Z}$. *Hint:* Use linearity of expectation and stationarity.

(b) Assuming $X \geq 0$, extend your argument to irrational $u > 0$. (The conclusion for general $X$ then follows by decomposing $X = X^+ - X^-$.)

*Solution.*

(a) We prove this for rationals. Let

$$m_t(u) = \mathbf{E}\left[\sum_{a \in A \cap (t, t+u]} X(a)\right].$$

Stationarity tells us that $m_0(u) = m_t(u)$ for all $t$ and all $u \geq 0$. Our goal is to show that $m_0$ is a linear function when restricted to rationals. Even without the restriction, it is additive by linearity of expectation and stationarity:

$$m_0(u + v) = m_0(u) + m_u(v) = m_0(u) + m_0(v).$$

So it remains only to show that $m_0(u) = u \cdot m_0(1)$. This holds if $u = 1/n$ for $n \in \mathbb{N}$ because

$$\frac{m_0(u)}{u} = n \cdot m_0(u) = \sum_{i=0}^{n} m_{iu}(u) = m_0(1),$$

where the last step uses linearity of expectation. Using this and additivity proves the result for all rationals.

(b) The short version is that $m_0$ is monotone when $X \geq 0$, which implies linearity over reals from linearity over rationals. To spell it out, let $u - \delta < u < u + \varepsilon$ be such that $u - \delta$ and $u + \varepsilon$ are rational. We can choose $\delta$ and $\varepsilon$ arbitrarily small. Then (a) implies

$$m_0(1) = \frac{m_0(u - \delta)}{u - \delta} = \frac{m_0(u + \varepsilon)}{u + \varepsilon}.$$

By monotonicity of $m_0$,

$$(u - \delta) \cdot m_0(1) \leq m_0(u) \leq (u + \varepsilon) \cdot m_0(1),$$

so the result follows by letting $\delta, \varepsilon \longrightarrow 0$.

**Exercise 1.2.** Let $A$ and $B$ be jointly stationary point processes that are (almost surely) disjoint.

(a) Show

$$\lambda_{A\cup B} = \lambda_A + \lambda_B.$$

(b) Show that for any process $X$ that is jointly stationary with $A$ and $B$,

$$\lambda_{A\cup B}\,\mathbf{E}_{A\cup B}[X] = \lambda_A\,\mathbf{E}_A[X] + \lambda_B\,\mathbf{E}_B[X].$$

This result is especially handy when applying RCL 1.22 to unions of point processes.

*Solution.*

(a) Definition 1.10 tells us that for a stationary point process $C$,

$$\lambda_C\,\mathbf{E}[\#(C \cap (0,1])].$$

Because $A$ and $B$ are disjoint,

$$\#((A\cup B)\cap(0,1]) = \#(A\cap(0,1]) + \#(A\cap(0,1]),$$

so by linearity of expectation,

$$\lambda_{A\cup B} = \mathbf{E}[\#((A\cup B)\cap(0,1])] = \mathbf{E}[\#(A\cap(0,1]) + \#(A\cap(0,1])] = \lambda_A + \lambda_B.$$

(b) Definition 1.20 tells us that for a point process $C$ jointly stationary with $X$,

$$\lambda_C\,\mathbf{E}_C[X] = \mathbf{E}\left[\sum_{t\in C\cap(0,1]} X(t)\right].$$

Because $A$ and $B$ are disjoint,

$$\sum_{t\in(A\cup B)\cap(0,1]} X(t) = \sum_{t\in A\cap(0,1]} X(t) + \sum_{t\in B\cap(0,1]} X(t),$$

so by linearity of expectation,

$$\begin{aligned}
\lambda_{A\cup B}\,\mathbf{E}_{A\cup B}[X] &= \mathbf{E}\left[\sum_{t\in(A\cup B)\cap(0,1]} X(t)\right] \\
&= \mathbf{E}\left[\sum_{t\in A\cap(0,1]} X(t) + \sum_{t\in B\cap(0,1]} X(t)\right] \\
&= \lambda_A\,\mathbf{E}_A[X] + \lambda_B\,\mathbf{E}_B[X].
\end{aligned}$$

The intuition here is that $\lambda_C\,\mathbf{E}_C[X]$ is in some sense an average accumulation rate. Imagine we keep track of a value, then add $X(t)$ to it at each time $t \in C$. Then $\lambda_C\,\mathbf{E}_C[X]$ is the average rate at which the value changes over time. This exercise shows that average rates of disjoint accumulations are additive.

**Exercise 1.3.** Let $A$ be a Poisson process. Explain why for any bounded function $f$,

$$\mathbf{E}[f(A \cup \{t\})] = \mathbf{E}_A^t[f(A)].$$

That is, explain why the distribution of $A \cup \{t\}$ under $\mathbf{P}[\cdot]$ is the same as the distribution of $A$ under $\mathbf{P}_A^t[\cdot]$. *Hint:* How would you generate a sample of the entire Poisson process $A$ "starting at" time $t$?

*Solution.* By Definition 1.14, both distributions are induced by the following simulation.

- Sample $R_i \sim \operatorname{Exp}(\lambda)$ freshly for all $i \in \mathbb{Z}$.
- Return the set

$$\{t\} \cup \left\{ t + \sum_{j=1}^{i} R_j : i \in \mathbb{N} \right\} \cup \left\{ t - \sum_{j=0}^{i} R_{-j} : i \in \mathbb{N} \right\}.$$

**Exercise 1.4.** Let $A$ be a stationary point process. Show

$$\lambda_A = \frac{1}{\mathbf{E}_A[A_1]}.$$

The intuition is that the rate $\lambda_A$ of the point process is the reciprocal of the average amount of time between its points, namely $\mathbf{E}_A[A_1] = \mathbf{E}_A[A_1 - A_0]$.

*Solution.* Applying PIF 1.25 to $X(t) \coloneqq 1$ yields

$$1 = \mathbf{E}[X] = \lambda_A \, \mathbf{E}_A\left[\int_0^{A_1} X(u) \, \mathrm{d}u\right] = \lambda_A \, \mathbf{E}_A[A_1].$$

**Exercise 1.5.** Let $W$ be a stationary M/G/1 work process. Show

$$\rho = \mathbf{P}[W > 0].$$

*Hint:* Use RCL 1.22, remembering the rule of thumb: consider, roughly, the integral of the function you're finding the expectation of. Thinking of $\mathbf{P}[W > 0] = \mathbf{E}[\mathbb{1}(W > 0)]$ as an expectation of a "zeroth-order" function of $W$, what does the rule of thumb suggest?

*Solution.* Applying RCL 1.22 to $W$ yields

$$0 = \mathbf{E}[\mathrm{D}W] + \lambda\, \mathbf{E}_A[\Delta W] = -\mathbf{E}[\mathbb{1}(W > 0)] + \lambda\, \mathbf{E}[S] = -\mathbf{P}[W > 0] + \rho.$$

**Exercise 1.6.** Let $X$ and $A$ be jointly stationary.

(a) Show

$$\mathbf{E}_A[X(A_1)] = \mathbf{E}_A[X(A_0)].$$

*Hint:* The right-hand side can be more simply written as $\mathbf{E}_A[X]$, because $A_0 = 0$ under $\mathbf{P}_A[\cdot]$ [Def. 1.20]. It's written the way it is as a suggestion of how you might use RCL 1.22 to prove it.

(b) Show that for all $i \in \mathbb{Z}$,

$$\mathbf{E}_A[X(A_i)] = \mathbf{E}_A[X(A_0)].$$

*Hint:* You can either adapt the argument you used for (a), or you can directly apply the result of (a) to $t \mapsto X(\text{something with } t)$.

*Solution.* We assume that $X$ is bounded for now. We'll explain how to extend to unbounded $X$ at the end.

(a) We apply RCL 1.22 to $X(A_1)$. The jump times are a subset of $A$, and the derivative is zero, so

$$0 = \mathbf{E}^t_A[X(A_1(t+)) - X(A_1(t-))] = \mathbf{E}^t_A[X(A_1(t)) - X(t)] = \mathbf{E}^t_A[X(A_1(t)) - X(A_0(t))].$$

(b) Applying (a) to $t \mapsto X(A_i(t))$ shows

$$\mathbf{E}_A[X(A_i)] = \mathbf{E}_A[X(A_i(A_0))] = \mathbf{E}_A[X(A_i(A_1))] = \mathbf{E}_A[X(A_{i+1})],$$

so the result follows by induction on $i$ (in both directions).

Suppose now that $X$ is unbounded. Let $i \in \mathbb{Z}$, and let $X_{m,n} = \min\{\max\{-m, X\}, n\}$. Then for all $m, n \geq 0$,

$$\mathbf{E}_A[X_{m,n}(A_i)] = \mathbf{E}_A[X_{m,n}(A_0)].$$

This is enough to tell us that if $\mathbf{E}_A[X(A_0)]$ is well defined, then so is $\mathbf{E}_A[X(A_i)]$, and the expectations are equal. In more detail:

- Letting $m = 0$ and taking $n \to \infty$ yields $\mathbf{E}_A[X(A_i)^+] = \mathbf{E}_A[X(A_0)^+]$.
- Letting $n = 0$ and taking $m \to \infty$ yields $\mathbf{E}_A[X(A_i)^-] = \mathbf{E}_A[X(A_0)^-]$.
- If one of these is infinite, then $\mathbf{E}_A[X(A_i)] = \mathbf{E}_A[X(A_0)]$ are both $\pm\infty$. Otherwise, $\mathbf{E}_A[X(A_i)] = \mathbf{E}_A[X(A_0)]$ are both finite.

**Exercise 1.7.** In this problem, you will explore variants of PIF 1.25. Let $X \geq 0$ and $A$ be jointly stationary.

(a) Show

$$\mathbf{E}[X] = \lambda_A \, \mathbf{E}_A\left[\int_{A_{-1}}^{0} X(u) \, \mathrm{d}u\right].$$

(b) Show

$$\mathbf{E}[X] = \lambda_A \, \mathbf{E}_A\left[\int_{\frac{1}{2}A_{-1}}^{\frac{1}{2}A_1} X(u) \, \mathrm{d}u\right].$$

*Solution.*

(a) The quick solution is to combine PIF 1.25 and Exercise 1.6. Letting

$$Y(t) := \int_{A_0(t)}^{A_1(t)} X(u) \, \mathrm{d}u$$

and noting $A_i(A_j) = A_{i+j}$, we want to show $\mathbf{E}[X] = \mathbf{E}_A[Y(A_{-1})]$. This follows because

- PIF 1.25 tells us $\mathbf{E}[X] = \mathbf{E}_A[Y] = \mathbf{E}_A[Y(A_0)]$, and
- Exercise 1.6 tells us $\mathbf{E}[Y(A_0)] = \mathbf{E}[Y(A_{-1})]$.

Here's a solution that doesn't rely on Exercise 1.6. Applying RCL 1.22 to

$$Z(t) := \int_{A_0(t)}^{t} X(u) \, \mathrm{d}u$$

and noting $A_0(t-) = A_{-1}(t)$ and $A_0(t) = t$ for $t \in A$ yields

$$\mathbf{E}[X(t)] = \mathbf{E}\big[(\mathrm{D}Z(t))^+\big] = \lambda_A \, \mathbf{E}_A^t\big[(\Delta Z(t))^-\big] = \lambda_A \, \mathbf{E}_A^t\left[\int_{A_{-1}(t)}^{t} X(u) \, \mathrm{d}u\right].$$

(b) A common trap for this part is applying RCL 1.22 to a process that isn't jointly stationary with $X$ and $A$, e.g. $t \mapsto X(t/2)$. Another trap was doing a change of integration variable: averaging the equations (a) and PIF 1.25 and substituting $v := u/2$ yields

$$\mathbf{E}[X] = \frac{\lambda_A}{2} \, \mathbf{E}_A\left[\int_{A_{-1}}^{A_1} X(u) \, \mathrm{d}u\right] = \lambda_A \, \mathbf{E}_A\left[\int_{\frac{1}{2}A_{-1}}^{\frac{1}{2}A_1} X(2v) \, \mathrm{d}v\right],$$

but we get a $2v$ where we really want just $v$.

One solution is to combine PIF 1.25 and Exercise 1.6, this time together with (a). Letting

$$Y_{\leq}(t) := \int_{A_0(t)}^{A_1(t)} X(u) \, \mathbb{1}\left(u \leq \frac{A_0(t) + A_1(t)}{2}\right) \mathrm{d}u,$$

$$Y_{>}(t) := \int_{A_0(t)}^{A_1(t)} X(u) \, \mathbb{1}\left(u > \frac{A_0(t) + A_1(t)}{2}\right) \mathrm{d}u$$

and noting $A_i(A_j) = A_{i+j}$, we want to show $\mathbf{E}[X] = \mathbf{E}_A[Y_{\leq}(A_0) + Y_{>}(A_{-1})]$. This follows because

- PIF 1.25 tells us $\mathbf{E}[X] = \mathbf{E}_A[Y_{\leq} + Y_{>}] = \mathbf{E}_A[Y_{\leq}(A_0) + Y_{>}(A_0)]$, and

- Exercise 1.6 tells us $\mathbf{E}[Y_>(A_0)] = \mathbf{E}[Y_>(A_{-1})]$.

  Here's another solution that doesn't rely on Exercise 1.6 or (a). We apply RCL 1.22 to

$$Y(t) \coloneqq \int_t^{\frac{1}{2}(A_0(t)+A_1(t))} X(u)\,\mathrm{d}u.$$

Note that integral's upper limit may be less than its lower limit, in which case the integral can be negative even though $X \geq 0$. Critically, $Y$ is jointly stationary with $X$ and $A$, because it can be written as

$$Y(t) = \int_t^{\frac{1}{2}((A_{\mathrm{shift}}(t))_0+(A_{\mathrm{shift}}(t))_1)} (X_{\mathrm{shift}}(t))(u)\,\mathrm{d}u.$$

See Remark 1.26 for further discussion. Applying RCL 1.22 $Y$ and reasoning through the values of $A_{-1}$, $A_0$, and $A_1$ right before and after time 0 under $\mathbf{P}_A[\cdot]$ (so $0 \in A$), we get

$$\mathbf{E}[X(t)] = \lambda_A\,\mathbf{E}_A\left[\int_0^{\frac{1}{2}(A_1+A_0)} X(u)\,\mathrm{d}u - \int_0^{\frac{1}{2}(A_0+A_{-1})} X(u)\,\mathrm{d}u\right] = \lambda_A\,\mathbf{E}_A\left[\int_{\frac{1}{2}A_{-1}}^{\frac{1}{2}A_1} X(u)\,\mathrm{d}u\right].$$

**Exercise 1.8.** Let $W$ be a stationary M/G/1 work process. Let a *(maximal) busy period* be a maximal contiguous interval of times $t$ during which $W(t) > 0$. Let $B$ and $C$ be the ends and starts, respectively, of busy periods. You may take as given that $B$ and $C$ are jointly stationary with $W$.

(a) Find $\lambda_C$, the average rate with which busy periods start. *Hint:* Think about the relationship between $C$ and the arrival times point process, then use PASTA 1.27.

(b) Find $\mathbf{E}_C[B_1]$, the mean length of a busy period. *Hint:* Use PIF 1.25.

*Hint:* You might find previous exercise helpful for both parts.

*Solution.* We know from Exercise 1.5 that $\mathbf{P}[W > 0] = \rho$, which is helpful for both parts. We also write $\lambda_A$ instead of $\lambda$ throughout to disambiguate it from $\lambda_C$.

(a) A busy period starts whenever a job arrives to an empty system. This happens at rate

$$\lambda_C = \lambda_A \mathbf{P}_A[W = 0],$$

where $A$ is the arrival times. (We give a more formal argument for this below.) Using PASTA 1.27, we get

$$\lambda_C = \lambda_A \mathbf{P}[W = 0] = \lambda_A(1 - \rho).$$

(b) We use PIF 1.25 on $\mathbb{1}(W > 0)$ and point process $C$. We know from Exercise 1.5 that $\mathbf{E}[\mathbb{1}(W > 0)] = \mathbf{P}[W > 0] = \rho$ (by a quick use of RCL 1.22 on $W$), so

$$\rho = \lambda_C \mathbf{E}_C\left[\int_0^{C_1} \mathbb{1}(W(u) > 0)\, \mathrm{d}u\right] = \lambda_C \mathbf{E}_C\left[\int_0^{B_1} 1\, \mathrm{d}u\right] = \lambda_C \mathbf{E}_C[B_1],$$

so

$$\mathbf{E}_C[B_1] = \frac{\rho}{\lambda_C} = \frac{\rho}{\lambda_A(1 - \rho)} = \frac{\mathbf{E}[S]}{1 - \rho}.$$

To argue $\lambda_C = \lambda_A \mathbf{P}_A[W = 0]$ more formally, we can use the definition of Palm expectation [Def. 1.20]. Letting

$$Z := \{t \in \mathbb{R} : W(t) = 0\}$$

be times when there is zero work, we have $C = A \cap Z$, from which we compute

$$\lambda_C = \mathbf{E}[N_C(0, 1]] = \mathbf{E}[\#(A \cap Z \cap (0, 1])] = \mathbf{E}\left[\sum_{a \in A \cap (0,1]} \mathbb{1}(W(a) = 0)\right] = \lambda_A \mathbf{P}_A[W = 0].$$

**Exercise 1.9** (Cavatappi). In this problem, you will prove an easy special case of PASTA 1.27 using PIF 1.25. Let $X \geq 0$ and $A$, a Poisson process, be jointly stationary and *independent*. Show

$$\mathbf{E}[X] = \mathbf{E}_A[X].$$

Specifically, use the fact that $A_1 \sim \operatorname{Exp}(\lambda_A)$ under $\mathbf{P}_A[\cdot]$ [Def. 1.14] to show

$$\lambda_A \, \mathbf{E}_A\left[\int_0^{A_1} X(u) \, \mathrm{d}u\right] = \mathbf{E}_A[X(A_1)],$$

then conclude using Exercise 1.6. *Hint:* It often helps to rewrite a random-domain integral as a deterministic-domain integral with an indicator in the integrand. Also, if $A$ and $X$ are independent, then $A$ and J$X$ are (almost surely) disjoint, so you can assume $X(a-) = X(a+)$ for all $a \in A$.

*Solution.* Using PIF 1.25, Definition 1.14, Tonelli's theorem, and the independence of $A_1$ and $X$ under $\mathbf{P}_A[\cdot]$, we compute

$$\begin{aligned}
\mathbf{E}[X] &= \lambda_A \, \mathbf{E}_A\left[\int_0^{A_1} X(u) \, \mathrm{d}u\right] \\
&= \lambda_A \int_0^\infty \mathbf{E}_A[X(u) \, \mathbb{1}(u < A_1)] \, \mathrm{d}u \\
&= \lambda_A \int_0^\infty \mathbf{E}_A[X(u)] \, \mathbf{P}[u < A_1] \, \mathrm{d}u \\
&= \int_0^\infty \mathbf{E}_A[X(u)] \, \lambda_A e^{-\lambda_A u} \, \mathrm{d}u \\
&= \int_0^\infty \mathbf{E}_A[X(A_1) \mid A_1 = u] \, \lambda_A e^{-\lambda_A u} \, \mathrm{d}u \\
&= \mathbf{E}_A[X(A_1)] \\
&= \mathbf{E}_A[X],
\end{aligned}$$

where the last step uses Exercise 1.6.

**Exercise 1.10** (Fettuccine). In this problem, you will prove a relatively easy, but still very useful, special case of PASTA 1.27 for the M/G/1. Let $W$ be a stationary M/G/1 work process, and let $f$ be a nonnegative function. Following the approach from Exercise 1.9, prove

$$\mathbf{E}[f(W)] = \mathbf{E}_A[f(W)],$$

where we recall that $W$ is left-continuous, so $W = W_{(-)}$ [Def. 1.13(c)]. You may take as given the fact that $W_{(+)}$ and $A_1$ are independent under $\mathbf{P}_A[\cdot]$. *Hint:* However, this does not imply that $W(u)$ is independent of $A_1$ under $\mathbf{P}_A[\cdot]$ for $u > 0$. Either argue why this extra independence holds, or come up with an approach that only needs independence of $W_{(+)} = W(0+)$ and $A_1$.

*Solution.* We can follow much the same strategy as the solution to Exercise 1.10. The key is that under $\mathbf{P}_A[\cdot]$ (so $0 = A_0 \in A$), for $u \in (0, A_1]$, we can express $W(u)$ in terms of the work $W = W(0)$ seen by the arrival at time 0, the size $S$ of the arrival, and the time $u$ since the arrival:

$$W(u) = (W + S - u)^+.$$

Using PIF 1.25, Definition 1.14, Tonelli's theorem, and the independence of $A_1$, $W$, and $S$ under $\mathbf{P}_A[\cdot]$, we compute

$$\begin{aligned}
\mathbf{E}[f(W)] &= \lambda_A \, \mathbf{E}_A\left[\int_0^{A_1} f(W(u)) \, \mathrm{d}u\right] \\
&= \lambda_A \int_0^\infty \mathbf{E}_A[f(W(u)) \, \mathbb{1}(u < A_1)] \, \mathrm{d}u \\
&= \lambda_A \int_0^\infty \mathbf{E}_A[f((W(0) + S - u)^+)] \, \mathbf{P}[u < A_1] \, \mathrm{d}u \\
&= \int_0^\infty \mathbf{E}_A[f((W + S - u)^+)] \, \lambda_A e^{-\lambda_A u} \, \mathrm{d}u \\
&= \int_0^\infty \mathbf{E}_A[f((W + S - A_1)^+) \mid A_1 = u] \, \lambda_A e^{-\lambda_A u} \, \mathrm{d}u \\
&= \mathbf{E}_A[f(W(A_1))] \\
&= \mathbf{E}_A[f(W)],
\end{aligned}$$

where the last step uses Exercise 1.6.

**Exercise 1.12.** Let $W$ be a stationary M/G/1 work process. Find a formula for $\mathbf{E}[e^{\theta W(t)}]$ using RCL 1.22 and PASTA 1.27.

(a) Do this assuming $\theta \le 0$. *Hint:* What, roughly, is the integral of $w \mapsto e^{\theta w}$?

(b) *Challenge!* Do this assuming $\theta > 0$, obtaining the same formula as in (a), but carefully tracking what preconditions are needed to ensure $\mathbf{E}[e^{\theta W(t)}] < \infty$. *Hint:* Apply RCL 1.22 to a truncated version of what you used in (a). You might get quantities that you can't analyze exactly, but you can bound them.

(c) *Open-ended....* Can you find the secret buses [§ 1.4.2] hiding your formula?

*Solution.*

(a) We follow the same strategy as the derivation of (1.3), except we apply RCL 1.22 to $e^{\theta W}$ instead of $W^2$. Because $e^{\theta W} \le 1$ for $\theta \le 0$, we satisfy the preconditions of RCL 1.22(b). Using PASTA 1.27 and the independence of $S$ and $W_{(-)} = W$ under $\mathbf{P}_A[\cdot]$ [Defs. 1.16, 1.13(c)], we compute

$$\begin{aligned} 0 &= \mathbf{E}[_t(e^{\theta W(t)})] + \lambda\,\mathbf{E}_A^t[\Delta_t(e^{\theta W(t)})] \\ &= \mathbf{E}[\theta e^{\theta W} \cdot \mathrm{D}W] + \lambda\,\mathbf{E}_A[e^{\theta(W_{(-)}+\Delta W)} - e^{\theta W_{(-)}}] \\ &= \theta\,\mathbf{E}[e^{\theta W}(-1 + \mathbb{1}(W = 0))] + \lambda\,\mathbf{E}[e^{\theta(W+S)} - e^{\theta W}] \\ &= -\theta\,\mathbf{E}[e^{\theta W}] + \theta\,\mathbf{P}[W = 0] + \lambda\,\mathbf{E}[e^{\theta W}]\,\mathbf{E}[e^{\theta S} - 1]. \end{aligned}$$

(Don't forget that $e^{\theta 0} = 1$, not 0.) Plugging in $\mathbf{P}[W = 0] = 1 - \rho$ [Exr. 1.5], we solve for

$$\mathbf{E}[e^{\theta W}] = \frac{\theta(1-\rho)}{\theta - \lambda\,\mathbf{E}[e^{\theta S} - 1]}.$$

(b) Roughly speaking, we're going to follow the same computation as in (a), but using RCL 1.22 on the truncated process $e^{\theta \min\{W,m\}}$, where $m > 0$ is a constant. This will yield a bound on $\mathbf{E}[e^{\theta W}\,\mathbb{1}(W \le m)]$, from which the same formula from (a) will follow by taking the $m \to \infty$ limit. We will find that the formula only holds for a limited range of $\theta$, e.g. $\mathbf{E}[e^{\theta S}] < \infty$ is necessary.

Before diving into the computation, it's worth considering: why should using RCL 1.22 on $e^{\theta \min\{W,m\}}$ yield a bound on $e^{\theta W}\,\mathbb{1}(W \le m)$? How does the minimum become an indicator? Remember the rule of thumb from Section 1.3.2: to get information about a process, apply RCL 1.22 to its "integral". Conversely, if we apply RCL 1.22 to a process, we should expect information about its "derivative". Indeed,

$$\mathrm{D}_t e^{\theta \min\{W(t),m\}} = \theta e^{\theta W(t)}\,\mathbb{1}(W(t) \le m) \cdot \mathrm{D}W(t).$$

For things to work out the way they did in (a), we would want the jump term to also result in $e^{\theta W(t)}\,\mathbb{1}(W(t) \le m)$ times something. However, for $t \in A$,

$$\Delta_t e^{\theta \min\{W(t),m\}} = e^{\theta W(t)}(e^{\theta \min\{S,(m-W(t))^+\}} - 1),$$

where we have a pesky $W(t)$ in the second factor on the right-hand side. The solution is to settle for an upper bound on the jump:

$$\Delta_t e^{\theta \min\{W(t),m\}} \leq e^{\theta W(t)}(e^{\theta S} - 1)\, \mathbb{1}(W(t) \leq m).$$

Here's the intuition behind this bound.

- If $W > m$, then $\min\{W, m\}$ doesn't change at all when we add $S$ to $W$.
- If $W \leq m$, then $\min\{W, m\}$ jumps up by at most $S$ when we add $S$ to $W$.

Another perspective on the bound is that $w \mapsto \min\{w, m\}$ is concave, and the bound on the jump term comes from the fact that a concave function is less than first-order approximations of it.

Having computed the derivative and jump terms, we can apply RCL 1.22 to $e^{\theta \min\{W,m\}}$, yielding

$$0 \leq -\theta\, \mathbf{E}[e^{\theta W}\, \mathbb{1}(W \leq m)] + \theta\, \mathbf{P}[W = 0] + \lambda\, \mathbf{E}[e^{\theta W}\, \mathbb{1}(W \leq m)]\, \mathbf{E}[e^{\theta S} - 1].$$

As long as

$$\lambda\, \mathbf{E}[e^{\theta S} - 1] < \theta,$$

then we can divide both sides of the inequality by $\theta - \lambda\, \mathbf{E}[e^{\theta S} - 1]$, obtaining

$$\mathbf{E}[e^{\theta W}\, \mathbb{1}(W \leq m)] \leq \frac{\theta(1-\rho)}{\theta - \lambda\, \mathbf{E}[e^{\theta S} - 1]}.$$

Taking $m \longrightarrow \infty$, the formula from (a) follows from monotone convergence theorem.

(c) Noticing that $\mathbf{E}[e^{\theta S} - 1]$ appears in Theorem 1.32(b), we might try rewriting $\mathbf{E}[e^{\theta W}]$ in terms of $\mathbf{E}[e^{\theta S_e}]$. Writing $\lambda = \rho/\mathbf{E}[S]$ to introduce the $1/\mathbf{E}[S]$ factor, this yields

$$\mathbf{E}[e^{\theta W}] = \frac{1-\rho}{1 - \rho\, \mathbf{E}[e^{\theta S_e}]}$$

under the condition

$$\rho\, \mathbf{E}[e^{\theta S_e}] < 1.$$

So there's a natural way to write $\mathbf{E}[e^{\theta W}]$ in terms of the excess $S_e$. But can we find the "geometrically many buses" [§ 1.4.2]? Well, because $\rho\, \mathbf{E}[e^{\theta S_e}] < 1$, we can write $\mathbf{E}[e^{\theta W}]$ as a geometric series:

$$\mathbf{E}[e^{\theta W}] = \sum_{n=0}^{\infty} (1-\rho)\rho^n\, \mathbf{E}[e^{\theta S_e}]^n.$$

Letting $N \sim \mathrm{Geo}_0(1-\rho)$ and $R_1, \ldots, R_N \sim S_e$ freshly as in Section 1.4.2, we have

$$\mathbf{E}[e^{\theta W}] = \sum_{n=0}^{\infty} \mathbf{P}[N = n]\, \mathbf{E}[e^{\theta S_e}]^n = \sum_{n=0}^{\infty} \mathbf{P}[N = n]\, \mathbf{E}\left[\exp\left(\theta \sum_{i=1}^{n} R_i\right)\right] = \mathbf{E}\left[\exp\left(\theta \sum_{i=1}^{N} R_i\right)\right].$$

This tells us the distribution of $W$ is the same as the distribution of the sum of $\mathrm{Geo}_0(1-\rho)$ many i.i.d. samples from $S_e$, i.e. "geometrically many buses".

Our finding here matches the heuristic "layers" argument [§ 1.4.2]. However, it doesn't yet precisely link the layers to the $S_e$ samples. This requires some more work, which you'll do in Exercise 1.18.

**Exercise 1.13.** The *M/G/1 with setup times (M/G/1/setup)* is a variant of the M/G/1, but with the following change: whenever a job arrives to an empty system, in addition to the job's size $S$ being added to the work, an additional *setup time*, sampled i.i.d. from a distribution $U$ on $[0, \infty)$, is also added. This represents the server taking extra time $U$ to set up after being idle.

Let $W$ be a stationary standard M/G/1/setup work process, and let $W_{\mathrm{M/G/1}}$ be a standard M/G/1 work process with the same arrival rate and size distribution.

(a) Find a formula for $\mathbf{E}[W]$ of the form

$$\mathbf{E}[W] = \mathbf{E}[W_{\mathrm{M/G/1}}] + \text{something},$$

where the M/G/1 and M/G/1/setup have the same arrival rate and size distribution.

(b) Can you interpret the "something" in your answer to (a) as the mean of some distribution? What might that distribution represent?

(c) Find a formula for $\mathbf{E}[e^{\theta W}]$. You should find a similar decomposition to what you found in (a). You may assume $\theta \leq 0$.

(d) Based on your answer to (c), was the distribution you found in (b) was correct, or did you find a different distribution that happens to have the right mean?

*Solution.*

(a) We follow the same strategy as the derivation of (1.3), with one change: when an arrival happens, we add $S + U$ instead of just $S$ if $W = 0$, where we recall that $W_{(-)} = W$ [Def. 1.13(c)]. One convenient way to visualize this is to write that under $\mathbf{P}_A[\cdot]$,

$$f(W_{(+)}) = f(W + S) + \mathbb{1}(W = 0)\,\big(f(S + U) - f(S)\big).$$

Applying RCL 1.22 to $W^2$ and using PASTA 1.27 and the independence assumptions of the M/G/1/setup, we compute

$$\begin{aligned}0 &= -2\,\mathbf{E}[W] + \lambda\,\mathbf{E}_A\big[(W + S)^2 - W^2 + \mathbb{1}(W = 0)\,\big((S + U)^2 - S^2\big)\big] \\ &= -2\,\mathbf{E}[W] + 2\lambda\,\mathbf{E}[S]\,\mathbf{E}[W] + \lambda\,\mathbf{E}[S^2] + \lambda\,\mathbf{P}[W = 0]\big(2\,\mathbf{E}[S]\,\mathbf{E}[U] + \mathbf{E}[U^2]\big),\end{aligned}$$

which, recalling $\rho \coloneqq \lambda\,\mathbf{E}[S]$ and Theorem 1.32(a), rearranges to

$$\mathbf{E}[W] = \frac{\rho\,\mathbf{E}[S_{\mathrm{e}}]}{1 - \rho} + \frac{\lambda\,\mathbf{P}[W = 0]\,\mathbf{E}[S]\,\mathbf{E}[U] + \frac{\lambda}{2}\,\mathbf{E}[U^2]}{1 - \rho}.$$

We can identify the first term as $\mathbf{E}[W_{\mathrm{M/G/1}}]$, so it remains to compute the second term.

The main task is to compute $\mathbf{P}[W = 0]$. Unlike in a standard M/G/1, *isn't* simply $1 - \rho$ [Exr. 1.5], because setup times change the work dynamics. But we can find $\mathbf{P}[W = 0]$ by applying RCL 1.22 to $W$. Once again using PASTA 1.27 and the definition of the M/G/1/setup, we compute

$$\begin{aligned}0 &= -1 + \mathbf{P}[W = 0] + \lambda\,\mathbf{E}_A[S + U\,\mathbb{1}(W = 0)] \\ &= -1 + \mathbf{P}[W = 0] + \rho + \lambda\,\mathbf{E}[U]\,\mathbf{P}[W = 0],\end{aligned}$$

so

$$\mathbf{P}[W = 0] = \frac{1 - \rho}{1 + \lambda\,\mathbf{E}[U]}.$$

Putting everything together, we get

$$\mathbf{E}[W] = \mathbf{E}[W_{\mathrm{M/G/1}}] + \frac{\lambda\,\mathbf{E}[U]}{1 + \lambda\,\mathbf{E}[U]}\,\mathbf{E}[U_{\mathrm{e}} + S].$$

(b) We can think of the extra term in (a) as being a probability times the mean of a positive random variable.

- The probability is $q := \lambda\,\mathbf{E}[U]/(1 + \lambda\,\mathbf{E}[U])$.
- The random variable is $U_{\mathrm{e}} + S$.

We interpret each of these in turn.

From our work in (a), we know $q = 1 - \mathbf{P}[W = 0]/(1 - \rho)$. To interpret this, the key idea is to think of $1 - \rho$ as a probability. We know from Exercise 1.5 that in a standard M/G/1, the server is busy with probability $\rho$. Even in the M/G/1/setup, it's still the case that the server is busy *with a job* (not a setup time) with probability $\rho$. (To show this formally, define a process which is the work from just jobs, excluding setup times, and apply RCL 1.22 to it.) This means

$$1 - \rho = \mathbf{P}[\text{server isn't busy with a job}],$$

and therefore

$$\frac{\mathbf{P}[W = 0]}{1 - \rho} = \mathbf{P}[\text{server isn't busy at all} \mid \text{server isn't busy with a job}],$$

so

$$q = 1 - \frac{\mathbf{P}[W = 0]}{1 - \rho} = \mathbf{P}[\text{server is busy with a setup time} \mid \text{server isn't busy with a job}].$$

Another way to see this directly from the definition of $q$ is to think about what $\lambda\,\mathbf{E}[U]$ and $1 + \lambda\,\mathbf{E}[U]$ might represent.

- $\lambda\,\mathbf{E}[U]$ is the average number of jobs that arrive during each setup time, *excluding* the job that triggers the start of the setup time.
- $1 + \lambda\,\mathbf{E}[U]$ is the average number of jobs that arrive during *or trigger the start of* each setup time.

So $q$ is the probability that a job arrives during a setup time, given that it either arrives during or triggers the start of a setup time. But these are exactly the jobs that arrive when the server isn't busy serving jobs.

So, if $q$ is a probability related to arriving during a setup time, how should we think about $U_{\mathrm{e}} + S$? Roughly speaking, we can think of it as an amount of "extra work" seen by arrivals that come during a setup time.

- The $U_{\mathrm{e}}$ could be the *remaining* setup time as observed by arrivals that occur during setup times [Rmk. 1.34, PASTA 1.27].

- The $S$ is not quite as clear, but one possibility is that it could be the work due to the job that started the setup time.

We might further guess that these are (conditionally) independent, but it isn't clear based on what we know so far.

(c) We can follow the same strategy as in (a), but applying RCL 1.22 to $e^{\theta W}$ instead of $W^2$. This yields a computation that looks much like the solution to Exercise 1.12(a), resulting in

$$0 = -\theta\,\mathbf{E}[e^{\theta W}] + \theta\,\mathbf{P}[W = 0] + \lambda\,\mathbf{E}[e^{\theta W}]\,\mathbf{E}[e^{\theta S} - 1] + \lambda\,\mathbf{P}[W = 0]\,\mathbf{E}[e^{\theta S}]\,\mathbf{E}[e^{\theta U} - 1].$$

Plugging in $\mathbf{P}[W = 0]$ from (a) and, inspired by (b), packaging some expressions using Theorem 1.32(b), we obtain

$$\begin{aligned}\mathbf{E}[e^{\theta W}] &= \frac{\mathbf{P}[W = 0]\bigl(1 + \lambda\,\mathbf{E}[U]\,\mathbf{E}[e^{\theta U_{\mathrm{e}}}]\,\mathbf{E}[e^{\theta S}]\bigr)}{1 - \rho\,\mathbf{E}[e^{\theta S_{\mathrm{e}}}]} \\ &= \frac{1 - \rho}{1 - \rho\,\mathbf{E}[e^{\theta S_{\mathrm{e}}}]} \cdot \frac{\lambda\,\mathbf{E}[U]}{1 + \lambda\,\mathbf{E}[U]}\,\mathbf{E}[e^{\theta U_{\mathrm{e}}}]\,\mathbf{E}[e^{\theta S}] \\ &= \mathbf{E}[e^{\theta W_{\mathrm{M/G/1}}}] \cdot \frac{\lambda\,\mathbf{E}[U]}{1 + \lambda\,\mathbf{E}[U]}\,\mathbf{E}[e^{\theta U_{\mathrm{e}}}]\,\mathbf{E}[e^{\theta S}].\end{aligned}$$

(d) The transform expression from (c) confirms our guess from (b). It tells us that the work in an M/G/1/setup is distributed as an independent sum of:

- The work $W_{\mathrm{M/G/1}}$ in a standard M/G/1 with the same M/G arrival process.
- A random variable distributed as

$$\begin{cases} U_{\mathrm{e}} + S & \text{with probability } \dfrac{\lambda\,\mathbf{E}[U]}{1 + \lambda\,\mathbf{E}[U]} \\ 0 & \text{otherwise,} \end{cases}$$

where the $U_{\mathrm{e}}$ and $S$ are independent.

**Exercise 1.14.** Repeat Exercise 1.13, but for the *M/G/1 with vacations (M/G/1/vacation)*. In the M/G/1/vacation, whenever the work reaches 0, it immediately jumps up by a *vacation* amount, which is sampled i.i.d from a distribution $V$ on $(0, \infty)$. This represents server taking a break whenever there's no work to do, coming back after time $V$.

The main difficulty of this problem is that in the M/G/1/vacation, $\Delta W$ is not simply the arrival times $A$. But you can partition $\Delta W = A \cup B$, where $B$ is the times when vacations start. One can show that $A$ and $B$ are (almost surely) disjoint, and you may use this fact without proof for this problem. *Hint:* You might find Exercise 1.2 handy. You can't use PASTA 1.27 under $\mathbf{P}_B[\cdot]$, so hopefully you won't need to. . . .

*Solution.* The reasoning in this problem is much like Exercises 1.12 and 1.13, so we focus most of the explanation below on handling the differences. For instance, we use PASTA 1.27 throughout without explicit mention.

There are two key observations to make about vacations.

- Under $\mathbf{P}_B[\cdot]$, we have $W = 0$, because vacations only start when the work reaches 0.
- Under $\mathbf{P}[\cdot]$, we have $W > 0$ and thus $W = -1$ almost surely, because vacations ensure that there is almost always work to do.

Throughout, we write the arrival rate as $\lambda_A = \lambda$ to disambiguate it from $\lambda_B$, the rate at which vacations start.

(a) We apply RCL 1.22 to $W^2$:

$$\begin{aligned} 0 &= -2\,\mathbf{E}[W] + \lambda_A\,\mathbf{E}_A[(W+S)^2 - W^2] + \lambda_B\,\mathbf{E}_B[(W+V)^2 - W^2] \\ &= -2\,\mathbf{E}[W] + 2\rho\,\mathbf{E}[W] + \lambda_A\,\mathbf{E}[S^2] + \lambda_B\,\mathbf{E}[V^2]. \end{aligned}$$

Using Theorem 1.32(a) to write as much as possible in terms of excess distributions yields

$$\mathbf{E}[W] = \frac{\rho\,\mathbf{E}[S_\mathrm{e}]}{1-\rho} + \frac{\lambda_B\,\mathbf{E}[V]}{1-\rho}\,\mathbf{E}[V_\mathrm{e}].$$

We recognize the first term as $\mathbf{E}[W_{\mathrm{M/G/1}}]$, so it remains to compute the second term.

The main task is to compute $\lambda_B$. How should we approach this? As usual, we apply RCL 1.22 to just the right process. In this case, $W$ does the trick. The intuition for why this should help is that $\lambda_B$ is related to the rate at which work increases due to vacations, and we already understand other ways work changes over time. Applying RCL 1.22 to $W$ yields

$$0 = -1 + \lambda_A\,\mathbf{E}_A[(W+S) - S]\lambda_B\,\mathbf{E}[(W+V) - V] = -1 + \rho + \lambda_B\,\mathbf{E}[V],$$

so

$$\lambda_B = \frac{1-\rho}{\mathbf{E}[V]}.$$

The intuition is that vacations need to fill the $1-\rho$ fraction of time that there isn't work from jobs [Exr. 1.5], and each vacation has average length $\mathbf{E}[V]$.

Putting everything together, we get

$$\mathbf{E}[W] = \mathbf{E}[W_{\mathrm{M/G/1}}] + \mathbf{E}[V_\mathrm{e}].$$

(b) The $V_\mathrm{e}$ could be the *remaining* vacation time as observed by arrivals that occur during vacations [Rmk. 1.34, PASTA 1.27].

(c) Applying RCL 1.22 to $e^{\theta W}$ yields

$$0 = -\theta\,\mathbf{E}[e^{\theta W}] + \lambda_A\,\mathbf{E}[e^{\theta W}]\,\mathbf{E}[e^{\theta S} - 1] + \lambda_B\,\mathbf{E}[e^{\theta V} - 1].$$

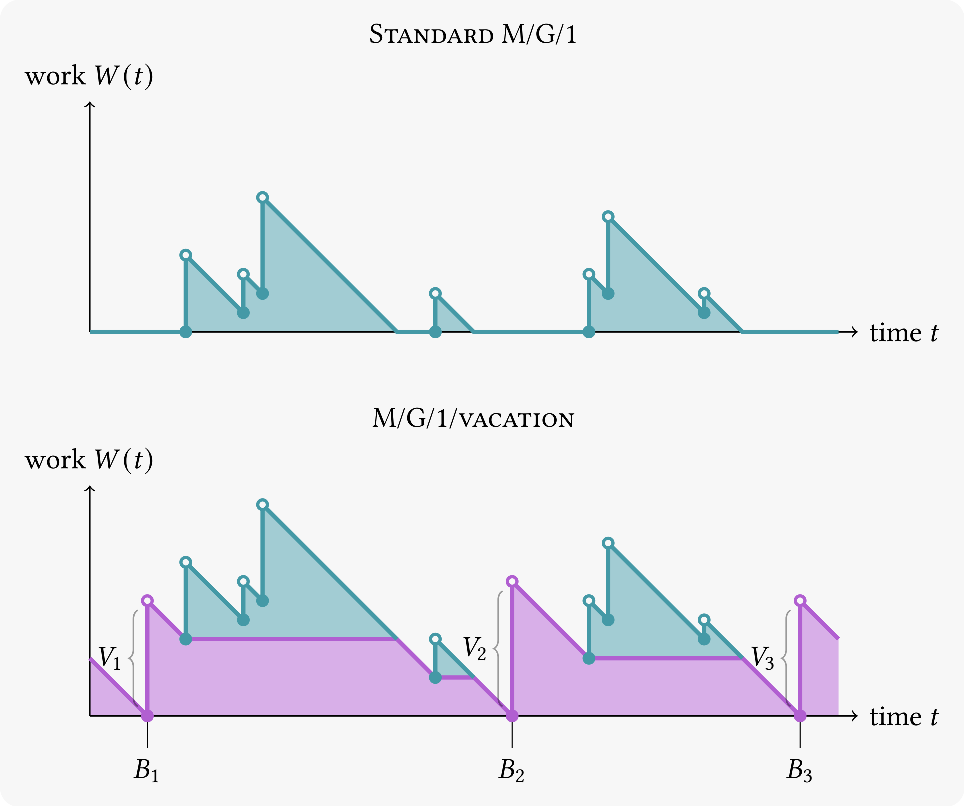
Plugging in the value of $\lambda_B$ from (a) and using Theorem 1.32(b) to write as much as possible in terms of excess distributions, this rearranges to

$$\mathbf{E}[e^{\theta W}] = \frac{1-\rho}{1-\rho\,\mathbf{E}[e^{\theta S_\mathrm{e}}]}\cdot\mathbf{E}[e^{\theta V_\mathrm{e}}] = \mathbf{E}[e^{\theta W_{\mathrm{M/G/1}}}]\cdot\mathbf{E}[e^{\theta V_\mathrm{e}}].$$

(d) The transform formula from (c) confirms our guess from (b). It tells us that the work in an M/G/1/setup is distributed as an independent sum of:

- The work $W_{\mathrm{M/G/1}}$ in a standard M/G/1 with the same M/G arrival process.
- A random variable with distribution $V_\mathrm{e}$.

There is an intuitive interpretation of this: along the same lines as Section 1.4.2 and Figure 1.3, we can view the extra work due to vacations as an extra "bottom layer" of work in the system, as illustrated below. The amount of work in the bottom layer has distribution $V_\mathrm{e}$ because it is, roughly speaking, the amount of work remaining in an in-progress vacation [Rmk. 1.34, PASTA 1.27]. This can be formalized using the strategy outlined in Exercise 1.18.



**Exercise 1.15.** *Challenge!* The *M/G/k* is a *multiserver* variant of the M/G/1. Specifically, let's imagine that the M/G/$k$ has $k$ "slow" servers, which run $k$ times slower than the single server of the M/G/1. A job of size $s$ thus takes time $ks$ to finish on one of the slow servers, but this is balanced out by the fact that there are $k$ servers. If all $k$ servers are busy at time $t$, then the M/G/$k$ is still completing work at rate 1, so $\mathrm{D}W(t) = -1$.

However, the M/G/$k$'s work process $W$ is *not standard* [Def. 1.13], because if there are fewer than $k$ jobs in the system at time $t$, then $\mathrm{D}W(t) \neq -1$, even if $W(t) > 0$. In fact, this reveals that $\mathrm{D}W(t)$ is no longer a deterministic function of $W(t)$: it depends on the number of jobs in the system, which we can't infer from $W(t)$ alone.

The above difficulties make analyzing the M/G/$k$'s mean work $\mathbf{E}[W]$ intractable in general. However, we can still get some useful formulas which lead to bounds under some conditions. The key idea is to define an *idleness process* [Def. 2.18]

$$I(t) := 1 - \frac{\text{\# jobs present at } t}{k} = \text{fraction of servers that are idle at } t,$$

then express $\mathrm{D}W(t)$ in terms of $I(t)$. You may assume $W$ and $I$ are jointly stationary.

(a) Show
$$\mathbf{E}[W] = \mathbf{E}[W_{\mathrm{M/G/1}}] + \frac{\mathbf{E}[IW]}{1-\rho}.$$

(b) Assuming there exists $m$ such that $\mathbf{P}[S \leq m] = 1$, show
$$\mathbf{E}[W] \leq \mathbf{E}[W_{\mathrm{M/G/1}}] + (k-1)m.$$
*Hint:* If $I(t) > 0$, how many jobs can there possibly be in the system at time $t$?

(c) Try to give an intuitive interpretation of the $\mathbf{E}[IW]/(1-\rho)$ term. *Hint:* Here's one somewhat heavy approach. Define the Palm-like expectation $\mathbf{E}_I[X] = \mathbf{E}[IX]/(1-\rho)$ for $X$ jointly stationary with $W$ and $I$. Just as $\mathbf{E}_A[\cdot]$ captures the perspective of an arriving job, consider: what perspective does $\mathbf{E}_I[\cdot]$ capture?

(d) Find a formula for $\mathbf{E}[e^{\theta W}]$ analogous to (a). You may assume $\theta \leq 0$. *Hint:* You should get $\mathbf{E}[e^{\theta W_{\mathrm{M/G/1}}}]$ [Exr. 1.12] times a factor that can be written using $\mathbf{E}_I[\cdot]$.

(e) *Open-ended....* To what extent do the above results generalize beyond the M/G/$k$?

*Solution.*

(a) This is a special case of Theorem 2.19.

(b) If $I > 0$, then at least one of the $k$ servers must be idle, so there are at most $k-1$ jobs in the system, each of which contributes at most $m$ work. Using RCL 1.22 on $W$ shows $\mathbf{E}[I] = 1 - \rho$ [Exr. 2.3], so

$$\frac{\mathbf{E}[IW]}{1-\rho} \leq \frac{\mathbf{E}[I \cdot (k-1)m]}{1-\rho} = (k-1)m.$$

# Bibliography

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

[2] Mor Harchol-Balter. 2013. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action.* Cambridge University Press, Cambridge, UK. doi:10.1017/CBO9781139226424.

[3] Mor Harchol-Balter, Mark E. Crovella, and Cristina D. Murta. 1999. On Choosing a Task Assignment Policy for a Distributed Server System. *J. Parallel and Distrib. Comput.* 59, 2 (Nov. 1999), 204–228. doi:10.1006/jpdc.1999.1577.

[4] David G. Kendall. 1953. Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of the Imbedded Markov Chain. *The Annals of Mathematical Statistics* 24, 3 (Sept. 1953), 338–354. doi:10.1214/aoms/1177728975.

[5] John F. C. Kingman. 1962. On Queues in Heavy Traffic. *Journal of the Royal Statistical Society: Series B (Methodological)* 24, 2 (July 1962), 383–392. doi:10.1111/j.2517-6161.1962.tb00465.x.

[6] Yuan Li and David A. Goldberg. 2024. Simple and Explicit Bounds for Multiserver Queues with $1/(1-\rho)$ Scaling. *Mathematics of Operations Research* (April 2024), moor.2022.0131. doi:10.1287/moor.2022.0131.

[7] Masakiyo Miyazawa. 1994. Rate Conservation Laws: A Survey. *Queueing Systems* 15, 1 (March 1994), 1–58. doi:10.1007/BF01189231.

[8] Ronald W. Wolff. 1982. Poisson Arrivals See Time Averages. *Operations Research* 30, 2 (April 1982), 223–231. doi:10.1287/opre.30.2.223.

[9] Runhan Xie, Isaac Grosof, and Ziv Scully. 2024. Heavy-Traffic Optimal Size- and State-Aware Dispatching. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 8, 1, Article 9 (March 2024), 36 pages. doi:10.1145/3639035.