Part IV

Computer Systems Modeling and Simulation

The goal of this part of the book is to learn how to run simulations of computer systems. Simulations are an important part of evaluating computer system performance. For example, we might have a new load-balancing algorithm, and we're trying to understand whether it reduces the mean job response time or improves utilization. Or we might have a queueing network, where we want to understand the fraction of packet drops when we double the arrival rate of packets. Being able to simulate the computer system is an easy way to get answers to such questions.

Before we can dive into the art of simulation, we first have to understand a few things about modeling. In Chapter 12 we study the Poisson process, which is the most common model used for the arrival process into a computer system. The Poisson process is not only easy to simulate, it also has many other beneficial properties when it comes to simulation and modeling.

In Chapter 13 we study the art of generating random variables for simulation. This is an extremely important part of simulation, since we often have to generate the interarrival times of jobs and the service requirements of jobs. Each of these is typically modeled by some random variable that is a good estimate of the empirical (true) workload. In our simulation, we need to generate instances of these random variables.

Finally, in Chapter 14 we are ready to understand how to program an event-driven simulation. We discuss several examples of event-driven simulation, focusing on the state that needs to be tracked and also on how to measure the quantities that we need from our simulation.

When simulating a computer system, we're often simulating a queueing network. We cover the basics of queueing networks in Chapter 14. However, we defer a more detailed discussion of queueing networks to Chapter 27, after we've covered Markov chains, which allow us to understand more about the analysis of queueing networks.

12 The Poisson Process

This chapter deals with one of the most important aspects of systems modeling, namely the *arrival process*. When we say "arrival process" we are referring to the sequence of arrivals into the system. The most widely used arrival process model is the Poisson process. This chapter defines the Poisson process and highlights its properties. Before we dive into the Poisson process, it will be helpful to review the Exponential distribution, which is closely related to the Poisson process.

12.1 Review of the Exponential Distribution

Recall we say that a random variable (r.v.) X is distributed Exponentially with *rate* λ, written $X \sim \text{Exp}(\lambda)$, if its probability density function (p.d.f.) is

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

The cumulative distribution function (c.d.f.), $F_X(x) = \mathbf{P}\{X \leq x\}$, is given by

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$\overline{F}_X(x) = 1 - F_X(x) = e^{-\lambda x}, \; x \geq 0.$$

Observe that both $f_X(x)$ and $\overline{F}_X(x)$ drop off by a *constant* factor, $e^{-\lambda}$, with each unit increase of x.

Recall also that for $X \sim \text{Exp}(\lambda)$, we have:

$$\mathbf{E}[X] = \frac{1}{\lambda} \qquad \mathbf{Var}(X) = \frac{1}{\lambda^2} \qquad C_X^2 = \frac{\mathbf{Var}(X)}{\mathbf{E}[X]^2} = 1.$$

In particular, the **rate** of the Exponential distribution, λ, is the reciprocal of its mean. Also recall that an Exponentially distributed r.v. X exhibits the **memoryless** property, which says that:

$$\mathbf{P}\{X > s + t \mid X > s\} = \mathbf{P}\{X > t\}, \quad \forall s, t \geq 0.$$

Finally, recall that the Exponential distribution has **constant failure rate** equal to λ (Exercise 10.2).

Question: Suppose that the lifetime of a job is Exponentially distributed with rate λ. Suppose that the job has already run for t seconds (its age is t). Consider a very small δ. What does the constant failure rate say about the probability that the job will complete in the next δ seconds?

Answer: The probability that a job of age t will complete in the next δ seconds is $\lambda\delta$, independent of t. See Chapter 10 for a review of the notion of failure rate.

12.2 Relating the Exponential Distribution to the Geometric

It can be proven that the Exponential distribution is the *only* continuous-time memoryless distribution.

Question: What is the only discrete-time memoryless distribution?

Answer: The Geometric distribution.

When reasoning about Exponential random variables, we find it very helpful to instead think about Geometric random variables, for which we have more intuition. We can think of the Exponential distribution as the "continuous counterpart" of the Geometric distribution by making the following analogy:

- The Geometric distribution can be viewed as the *number* of flips needed to get a "success." The distribution of the remaining *number* of flips is independent of how many times we have flipped so far.
- The Exponential distribution is the *time* until "success." The distribution of the remaining *time* is independent of how long we have waited so far.

To unify the Geometric and Exponential distributions, we introduce the notion of a "**δ-step proof**." Throughout the chapter, we will use this way of thinking to come up with quick intuitions and arguments. The idea is to imagine each unit of time as divided into n pieces, each of duration $\delta = \frac{1}{n}$, and suppose that a trial (coin flip) occurs every δ time period, rather than at unit times.

We now define a r.v. Y, where Y is Geometrically distributed with probability $p = \lambda\delta$ of getting a head, for some small $\delta \to 0$. However, rather than flipping every unit time step, we flip every δ-step. That is,

$$Y \sim \text{Geometric}(p = \lambda\delta \mid \text{ Flip every } \delta\text{-step}).$$

Observe that Y denotes the *number* of flips until success. Now define Y^* to be the *time* until success under Y:

$$Y^* = \text{Time associated with } Y.$$

Observe that as $\delta \to 0$ (or $n \to \infty$), Y^* becomes a positive, real-valued r.v., because success can occur at any time.

Question: What is $\mathbf{E}[Y^*]$? How is Y^* distributed?

Answer:

$$\begin{aligned}\mathbf{E}[Y^*] &= (\text{avg. \# trials until success}) \cdot (\text{time per trial}) \\ &= \frac{1}{\delta\lambda} \cdot \delta = \frac{1}{\lambda}.\end{aligned}$$

To understand the distribution of Y^*, we express $\mathbf{P}\{Y^* > t\}$ as the probability that all the trials up to at least time t have been failures (i.e., we have had at least t/δ failures).

$$\begin{aligned}\mathbf{P}\{Y^* > t\} = \mathbf{P}\left\{\text{at least } \frac{t}{\delta} \text{ failures}\right\} &= (1-\delta\lambda)^{\frac{t}{\delta}} \\ &= \left(1 - \frac{1}{\frac{1}{\delta\lambda}}\right)^{\frac{t}{\delta}} \\ &= \left(1 - \frac{1}{\frac{1}{\delta\lambda}}\right)^{\frac{1}{\lambda\delta}\cdot\lambda\cdot t} \\ &= \left[\left(1 - \frac{1}{\frac{1}{\delta\lambda}}\right)^{\frac{1}{\lambda\delta}}\right]^{\lambda t} \\ &\longrightarrow [e^{-1}]^{\lambda t}, \text{ as } \delta \to 0, \text{ by (1.9)} \\ &= e^{-\lambda t}.\end{aligned}$$

But $\mathbf{P}\{Y^* > t\} = e^{-\lambda t}$ implies that $Y^* \sim \text{Exp}(\lambda)$.

We have thus proven the following theorem, which is depicted in Figure 12.1.

Theorem 12.1 *Let $X \sim Exp(\lambda)$. Then X represents the time to a successful event, given that an event occurs every δ-step and is successful with probability $\lambda\delta$, where $\delta \to 0$.*

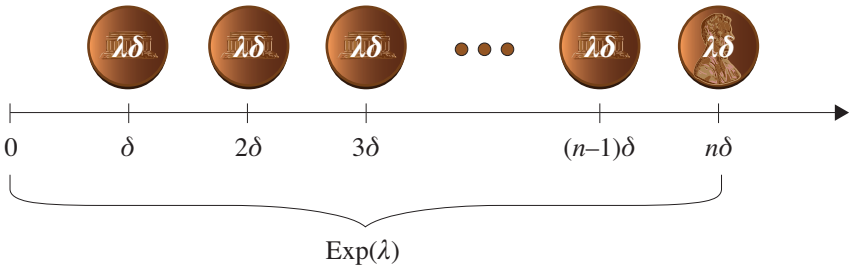

Figure 12.1 *Geometric depiction of the Exp(λ) distribution. Time is divided into steps of duration δ, and a coin (with probability $\lambda\delta$ of "heads") is flipped only at each δ-step.*

12.3 More Properties of the Exponential

Before we continue, here is a useful definition.

Definition 12.2

$$f = o(\delta) \quad \text{if} \quad \lim_{\delta \to 0} \frac{f}{\delta} = 0.$$

For example, $f = \delta^2$ is $o(\delta)$ because $\frac{\delta^2}{\delta} \to 0$ as $\delta \to 0$. Likewise $f = \sqrt{\delta}$ is *not* $o(\delta)$. Basically, a function is $o(\delta)$ if it goes to zero faster than δ, as $\delta \to 0$.

This definition may seem a little odd, because in general asymptotic notation (as in Section 1.6) "big-O" and "little-o" are defined in terms of some $n \to \infty$, not as $\delta \to 0$. When we use $\delta \to 0$, everything is flipped.

We now illustrate how to combine the $o(\delta)$ notation with the discretized view of an Exponential to prove a few properties of the Exponential distribution.

Theorem 12.3 *Given $X_1 \sim Exp(\lambda_1)$, $X_2 \sim Exp(\lambda_2)$, $X_1 \perp X_2$,*

$$\mathbf{P}\{X_1 < X_2\} = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

Proof: **(Traditional algebraic proof)**

$$\begin{aligned}
\mathbf{P}\{X_1 < X_2\} &= \int_0^\infty \mathbf{P}\{X_1 < X_2 \mid X_2 = x\} \cdot f_2(x)dx \\
&= \int_0^\infty \mathbf{P}\{X_1 < x \mid X_2 = x\} \cdot \lambda_2 e^{-\lambda_2 x} dx \\
&= \int_0^\infty \mathbf{P}\{X_1 < x\} \cdot \lambda_2 e^{-\lambda_2 x} dx, \qquad \text{since } X_1 \perp X_2
\end{aligned}$$

Continuing,

$$\begin{aligned}\mathbf{P}\{X_1 < X_2\} &= \int_0^\infty (1 - e^{-\lambda_1 x})(\lambda_2 e^{-\lambda_2 x})dx \\ &= \int_0^\infty \lambda_2 e^{-\lambda_2 x}dx - \lambda_2 \int_0^\infty e^{-(\lambda_1+\lambda_2)x}dx \\ &= 1 - \frac{\lambda_2}{\lambda_1+\lambda_2} \\ &= \frac{\lambda_1}{\lambda_1+\lambda_2}.\end{aligned}$$

■

Now for a more intuitive proof, by analogy with the Geometric distribution:

Proof: **(Intuitive Geometric proof)** Success of type 1 occurs with probability $\lambda_1\delta$ on each δ-step. Independently, success of type 2 occurs with probability $\lambda_2\delta$ on each δ-step. $\mathbf{P}\{X_1 < X_2\}$ is really asking, given that a success of type 1 or type 2 has occurred, what is the probability that it is a success of type 1?

$$\begin{aligned}\mathbf{P}\{\text{type 1} \mid \text{type 1 or type 2}\} &= \frac{\mathbf{P}\{\text{type 1}\}}{\mathbf{P}\{\text{type 1 or type 2}\}} \\ &= \frac{\lambda_1\delta}{\lambda_1\delta + \lambda_2\delta - (\lambda_1\delta)(\lambda_2\delta)} \\ &= \frac{\lambda_1\delta}{\lambda_1\delta + \lambda_2\delta - o(\delta)} \\ &= \frac{\lambda_1}{\lambda_1 + \lambda_2 - \frac{o(\delta)}{\delta}} \\ &\to \frac{\lambda_1}{\lambda_1+\lambda_2} \text{ as } \delta \to 0.\end{aligned}$$

■

Example 12.4 (Which fails first?)

There are two potential failure points for our server: the power supply and the disk. The lifetime of the power supply is Exponentially distributed with mean 500, and the lifetime of the disk is independently Exponentially distributed with mean 1,000.

Question: What is the probability that the system failure, when it occurs, is caused by the power supply?

Answer: $\frac{\frac{1}{500}}{\frac{1}{500}+\frac{1}{1000}}$.

Theorem 12.5 *Given* $X_1 \sim Exp(\lambda_1)$, $X_2 \sim Exp(\lambda_2)$, $X_1 \perp X_2$. *Let*

$$X = \min(X_1, X_2).$$

Then

$$X \sim Exp(\lambda_1 + \lambda_2).$$

Proof: **(Traditional algebraic proof)**

$$\begin{aligned}
\mathbf{P}\{X > t\} &= \mathbf{P}\{\min(X_1, X_2) > t\} \\
&= \mathbf{P}\{X_1 > t \text{ and } X_2 > t\} \\
&= \mathbf{P}\{X_1 > t\} \cdot \mathbf{P}\{X_2 > t\} \\
&= e^{-\lambda_1 t} \cdot e^{-\lambda_2 t} \\
&= e^{-(\lambda_1+\lambda_2)t}.
\end{aligned}$$

■

Here is an alternative argument by analogy with the Geometric distribution:

Proof: **(Intuitive Geometric proof)**

- A trial occurs every δ-step.
- The trial is "successful of type 1" with probability $\lambda_1\delta$.
- The trial is "successful of type 2" independently with probability $\lambda_2\delta$.
- We are looking for the time until there is a success of either type.
  A trial is "successful" (either type) with probability

$$\lambda_1\delta + \lambda_2\delta - (\lambda_1\delta)\cdot(\lambda_2\delta) = \delta \underbrace{\left(\lambda_1 + \lambda_2 - \frac{o(\delta)}{\delta}\right)}_{\text{rate}}.$$

- Thus the time until we get a "success" is Exponentially distributed with rate

$$\lambda_1 + \lambda_2 - \frac{o(\delta)}{\delta},$$

  and as $\delta \to 0$ this gives the desired result. ■

Question: In the server from Example 12.4, what is the time until there is a failure of either the power supply or the disk?

Answer: Exponential with rate $\left(\frac{1}{500} + \frac{1}{1000}\right)$.

12.4 The Celebrated Poisson Process

The Poisson process is the most widely used model for arrivals into a system. Part of the reason for this is that it is analytically tractable. However, the Poisson process is also a good model for any process of arrivals which is the aggregation of many independently behaving users. For example, the Poisson process is a good representation of the arrivals of requests into a web server, or the arrivals of jobs into a supercomputing center, or the arrivals of emails into a mail server. The "Limiting Theorem," see [45, pp. 221–228] explains how an aggregate of independent arrival processes leads to a Poisson process. The point is this: If you look at the request stream from an individual user, it will *not* look like a Poisson process. However, if you aggregate the requests from a very large number of users, that *aggregate stream* starts to look like a Poisson process.

Before we define a Poisson process, it helps to recall the Poisson distribution.

Question: If $X \sim \text{Poisson}(\lambda)$, what is $p_X(i)$, $\mathbf{E}[X]$, and $\mathbf{Var}(X)$?

Answer:

$$p_X(i) = \frac{e^{-\lambda}\lambda^i}{i!}, \quad i = 0, 1, 2, \ldots$$

$$\mathbf{E}[X] = \mathbf{Var}(X) = \lambda.$$

A Poisson process is a particular type of arrival sequence. We will need a little terminology. Figure 12.2 shows a sequence of arrivals. Each arrival is associated with a time. The arrival times are called "events."

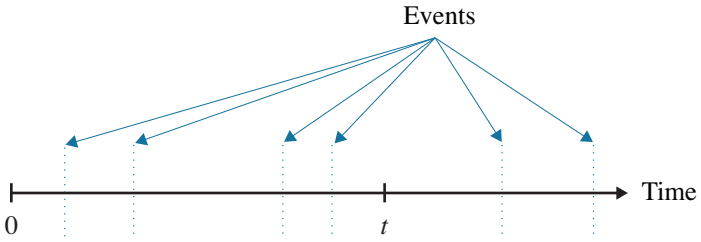

Figure 12.2 *Sequence of events.*

Definition 12.6 *For any sequence of events, we define $N(t)$, $t \geq 0$ to be the* **number of events that occurred by time *t*** *(including time t).*

Definition 12.7 *An event sequence has* **independent increments** *if the numbers of events that occur in disjoint time intervals are independent. Specifically, for all $t_0 < t_1 < t_2 < \ldots < t_n$, the n quantities below are independent:*

$$N(t_1) - N(t_0) \perp N(t_2) - N(t_1) \perp \ldots \perp N(t_n) - N(t_{n-1}).$$

Example 12.8 (Examples of sequences of events)

Consider three sequences of events:

(a) births of children
(b) people entering a store
(c) goals scored by a particular soccer player.

Question: Do these event processes have independent increments?

Answer:

(a) No. The number of births depends on the population size, which increases with prior births.
(b) Yes.
(c) Maybe. Depends on whether we believe in slumps!

Definition 12.9 *The event sequence has* **stationary increments** *if the number of events during a time period depends only on the length of the time period and not on its starting point. That is, $N(t+s) - N(s)$ has the same distribution for all s.*

Definition 12.10 (First definition of the Poisson process) *A Poisson process with rate λ is a sequence of events such that*

1. *$N(0) = 0$.*
2. *The process has independent increments.*
3. *The number of events in any interval of length t is Poisson distributed with mean λt. That is, $\forall s, t \geq 0$,*

$$\mathbf{P}\{N(t+s) - N(s) = n\} = \frac{e^{-\lambda t}(\lambda t)^n}{n!} \quad n = 0, 1, \ldots$$

Question: Why is λ called the "rate" of the process?

Answer: Observe that $\mathbf{E}[N(t)] = \lambda t$, so the rate of events is $\frac{\mathbf{E}[N(t)]}{t} = \lambda$.

Question: Why only "independent increments"?

Answer: The third item in the definition already implies stationary increments, because the number of events within an interval of length t depends only on t.

Observe that the assumption of stationary and independent increments is equivalent to asserting that, at any point in time, the process *probabilistically restarts itself*. That is, the process from any point on is independent of all that occurred previously (by independent increments) and also has the same distribution as the original process (by stationary increments). Simply put, the process has no memory. This leads us to the second definition of the Poisson process.

Definition 12.11 (Second definition of the Poisson process) *A Poisson process with rate λ is a sequence of events such that the inter-event times are i.i.d. Exponential random variables with rate λ and $N(0) = 0$.*

Question: Which definition of a Poisson process would you use when trying to simulate a Poisson process, the first or the second?

Answer: The Second Definition seems much easier to work with. The times between arrivals are just instances of $\text{Exp}(\lambda)$. We will learn how to generate instances of $\text{Exp}(\lambda)$ in Chapter 13.

First Definition $\Rightarrow$ Second Definition

Let $T_1, T_2, T_3, \ldots$ be the *inter-event* times of a sequence of events. We need to show that $T_i \sim \text{Exp}(\lambda)$, $\forall i$. By the first definition,

$$\mathbf{P}\{T_1 > t\} = \mathbf{P}\{N(t) = 0\} = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t}.$$

Next,

$$\begin{aligned}\mathbf{P}\left\{T_{n+1} > t \,\middle|\, \sum_{i=1}^{n} T_i = s\right\} &= \mathbf{P}\left\{0 \text{ events in } (s, s+t) \,\middle|\, \sum_{i=1}^{n} T_i = s\right\} \\ &= \mathbf{P}\{0 \text{ events in } (s, s+t)\}, \quad \text{by indpt. increments} \\ &= e^{-\lambda t}, \quad \text{by stationary increments.}\end{aligned}$$

Second Definition $\Rightarrow$ First Definition

Feller [27, p. 11] has a rigorous algebraic proof that the Second Definition implies the First Definition. The idea is to show that the sum of n i.i.d. $\text{Exp}(\lambda)$

random variables has a Gamma, $\Gamma(n, \lambda)$ distribution. Feller then uses the $\Gamma(n, \lambda)$ distribution to show that $N(t)$ follows a Poisson distribution.

Rather than going through this tedious algebraic proof, we instead provide an argument by analogy with the Geometric distribution: $N(t)$ refers to the number of arrivals by time t. Our goal is to prove that $N(t) \sim \text{Poisson}(\lambda t)$. Think of an arrival/event as being a "success." The fact that the interarrival times are distributed as $\text{Exp}(\lambda)$ corresponds to flipping a coin every δ-step, where a flip is a success (arrival) with probability $\lambda\delta$:

$$\begin{aligned} N(t) &= \text{Number of successes (arrivals) by time } t \\ &\sim \text{Binomial}(\#\text{ flips, probability of success of each flip}) \\ &\sim \text{Binomial}\left(\frac{t}{\delta}, \lambda\delta\right). \end{aligned}$$

Observe that as $\delta \to 0$, $\frac{t}{\delta}$ becomes very large and $\lambda\delta$ becomes very small.

Question: Now what do you know about $\text{Binomial}(n, p)$ for large n and tiny p?

Answer: Recall from Exercise 3.8 that

$$\text{Binomial}(n, p) \to \text{Poisson}(np), \quad \text{as } n \to \infty \text{ and } p \to 0.$$

So, as $\delta \to 0$,

$$N(t) \sim \text{Poisson}\left(\frac{t}{\delta} \cdot \lambda\delta\right) = \text{Poisson}(\lambda t).$$

12.5 Number of Poisson Arrivals during a Random Time

Imagine that jobs arrive to a system according to a Poisson process with rate λ. We wish to understand how many arrivals occur during time S, where S is a r.v. Here, S might represent the time that a job is being processed. Assume that S is independent of the Poisson process. Let A_S denote the number of Poisson arrivals during S. It is useful to first talk about A_t, the number of arrivals during a constant time t. Notice that A_t is what we normally refer to as $N(t)$.

Definition 12.12 *Assume that arrivals occur according to a Poisson process with rate λ. We define*

$$A_t = N(t) = \textit{Number of arrivals during time } t$$

and

$$A_S = \textit{Number of arrivals during time r.v. } S.$$

Question: What is $\mathbf{E}[A_t]$?

Answer: $\mathbf{E}[A_t] = \mathbf{E}[N(t)] = \lambda t$.

Question: What is $\mathbf{Var}(A_t)$?

Answer: Recall that $A_t = N(t) \sim \text{Poisson}(\lambda t)$. Thus $\mathbf{Var}(A_t) = \lambda t$.

Question: If we want to know the moments of A_S, what should we do?

Answer: Condition on the value of S. For example, to get the first moment of A_S we write:

$$\begin{aligned}\mathbf{E}[A_S] &= \int_{t=0}^{\infty} \mathbf{E}[A_S \mid S = t] \cdot f_S(t)dt \\ &= \int_{t=0}^{\infty} \mathbf{E}[A_t] \cdot f_S(t)dt \\ &= \int_{t=0}^{\infty} \lambda t \cdot f_S(t)dt \\ &= \lambda \mathbf{E}[S]. \end{aligned} \tag{12.1}$$

12.6 Merging Independent Poisson Processes

In networks, it is common that two Poisson processes are *merged*, meaning that they're interleaved into a single process as shown in Figure 12.3.

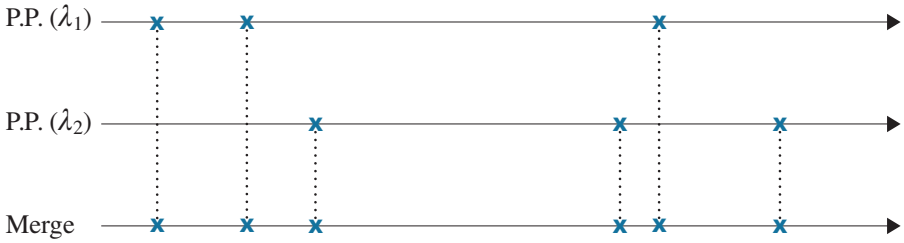

Figure 12.3 *A Poisson process with rate λ_1 is merged with a Poisson process with rate λ_2.*

Theorem 12.13 (Poisson merging) *Given two independent Poisson processes, where process 1 has rate λ_1 and process 2 has rate λ_2, the merge of process 1 and process 2 is a single Poisson process with rate $\lambda_1 + \lambda_2$.*

Proof: Process 1 has $\text{Exp}(\lambda_1)$ interarrival times. Process 2 has $\text{Exp}(\lambda_2)$ inter-

arrival times. The time until the first event from either process 1 or process 2 is the minimum of $\text{Exp}(\lambda_1)$ and $\text{Exp}(\lambda_2)$, which is distributed $\text{Exp}(\lambda_1 + \lambda_2)$ (Theorem 12.5). Likewise, the time until the second event is also distributed $\text{Exp}(\lambda_1 + \lambda_2)$, etc. Thus, using the Second Definition, we have a Poisson process with rate $\lambda_1 + \lambda_2$. ■

Proof: **(Alternative)** Let $N_i(t)$ denote the number of events in process i by time t:

$$N_1(t) \sim \text{Poisson}(\lambda_1 t)$$
$$N_2(t) \sim \text{Poisson}(\lambda_2 t).$$

Yet the sum of two independent Poisson random variables is still Poisson with the sum of the means, so

$$\underbrace{N_1(t) + N_2(t)}_{\text{merged process}} \sim \text{Poisson}(\lambda_1 t + \lambda_2 t).$$

■

12.7 Poisson Splitting

It is also common that a stream of arrivals is split into two streams, where each arrival is sent to the A stream with probability p and to the B stream with probability $1 - p$. Figure 12.4 illustrates the splitting of a Poisson stream.

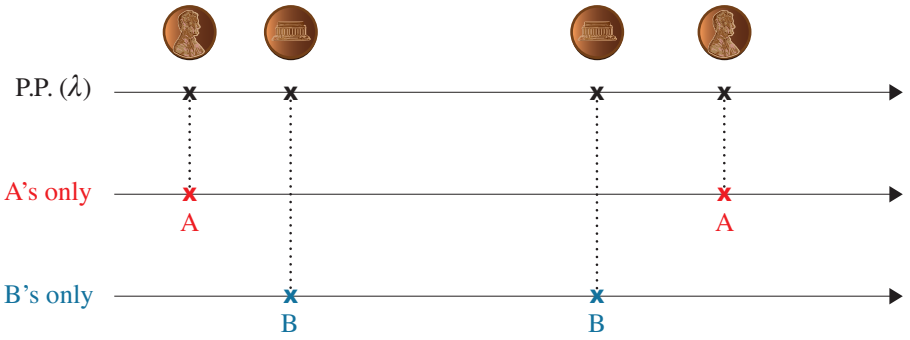

Figure 12.4 *Splitting a Poisson process with rate λ into an A stream and a B stream, based on coin flips.*

Theorem 12.14 (Poisson splitting) *Given a Poisson process with rate λ, suppose that each event is classified "type A" with probability p and "type B" with probability $1-p$. Then type A events form a Poisson process with rate $p\lambda$, type B events form a Poisson process with rate $(1-p)\lambda$, and these two processes are independent. Specifically, if $N_A(t)$ denotes the number of type A events by time t, and $N_B(t)$ denotes the number of type B events by time t, then*

$$\begin{aligned}\mathbf{P}\{N_A(t)=n, N_B(t)=m\} &= \mathbf{P}\{N_A(t)=n\}\cdot\mathbf{P}\{N_B(t)=m\}\\ &= e^{-\lambda t p}\frac{(\lambda t p)^n}{n!}\cdot e^{-\lambda t(1-p)}\frac{(\lambda t(1-p))^m}{m!}.\end{aligned}$$

This is one of those theorems that initially seems very counter-intuitive. It is really not clear why the times between the type A events end up being *Exponentially* distributed with rate λp as opposed to something else. Consider the sequence of events comprising the original Poisson process, where a coin with bias p is flipped at each event. When the coin flip comes up "head," the event is classified as "type A." If we look at just the type A events, we might imagine that some pairs of consecutive type A events are separated by $\text{Exp}(\lambda)$ (where we had two heads in a row) while other pairs of consecutive type A events are separated by multiple $\text{Exp}(\lambda)$ periods (where we didn't have a head for a while). It is not at all clear why the times between type A events are actually $\text{Exp}(\lambda p)$.

Before proving Theorem 12.14, we provide intuition for what's going on, by again making use of δ-step arguments. The original process has $\text{Exp}(\lambda)$ interarrival times, which is equivalent to tossing a coin every $\delta \to 0$ steps, where the coin comes up "success" with probability $\lambda\delta$. We refer to this $\lambda\delta$ coin as the *first* coin. Now we can imagine a *second* coin being flipped, where the second coin has probability p of success. Only if *both* the first and second coins are successes at the same time do we have a type A success. But this is equivalent to flipping just a *single* coin, with probability $\lambda\delta p$ of success. The time between successes for the single coin is then distributed $\text{Exp}(\lambda p)$. This proof is illustrated in Figure 12.5 and can be repeated for type B events.

Proof: [Theorem 12.14] This proof is taken from [64, p. 258]. What makes this proof precise is that (1) it uses no approximations and (2) it explicitly proves independence. Let

$$\begin{aligned}N(t) &= \text{Number of events by time } t \text{ in the original process}\\ N_A(t) &= \text{Number of type A events by time } t\\ N_B(t) &= \text{Number of type B events by time } t.\end{aligned}$$

We start by computing the joint probability that there are n events of type A and

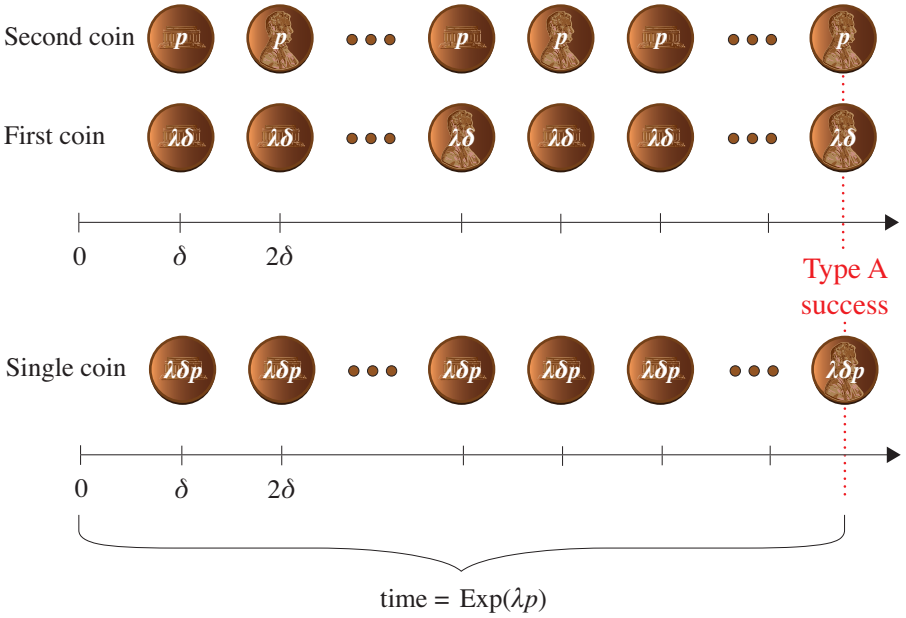

Figure 12.5 *A "type A success" only occurs if both the $\lambda\delta$-coin and the p-coin are heads.*

m events of type B by time t.

$$\begin{aligned}
&\mathbf{P}\{N_A(t) = n, N_B(t) = m\} \\
&= \sum_{k=0}^{\infty} \mathbf{P}\{N_A(t) = n, N_B(t) = m \mid N(t) = k\} \cdot \mathbf{P}\{N(t) = k\} \\
&= \mathbf{P}\{N_A(t) = n, N_B(t) = m \mid N(t) = n+m\} \cdot \mathbf{P}\{N(t) = n+m\} \\
&\quad \text{(because this is the only non-zero term in the above sum)} \\
&= \mathbf{P}\{N_A(t) = n, N_B(t) = m \mid N(t) = n+m\} \cdot e^{-\lambda t} \frac{(\lambda t)^{n+m}}{(n+m)!} \\
&= \binom{n+m}{n} p^n (1-p)^m e^{-\lambda t} \frac{(\lambda t)^{n+m}}{(n+m)!},
\end{aligned}$$

where the last line comes from the Binomial.

Simplifying, we have:

$$\begin{aligned}
\mathbf{P}\{N_A(t) = n, N_B(t) = m\} &= \frac{(m+n)!}{n!m!} p^n (1-p)^m e^{-\lambda t} \frac{(\lambda t)^{n+m}}{(n+m)!} \\
&= e^{-\lambda t p} \frac{(\lambda t p)^n}{n!} \cdot e^{-\lambda t(1-p)} \frac{(\lambda t(1-p))^m}{m!}. \quad (12.2)
\end{aligned}$$

To illustrate that the type A process and type B process are independent, we now compute the marginal probability $\mathbf{P}\{N_A(t) = n\}$ by summing the joint

probability, (12.2), over all values of m:

$$\begin{aligned}\mathbf{P}\{N_A(t)=n\} &= \sum_{m=0}^{\infty} \mathbf{P}\{N_A(t)=n, N_B(t)=m\} \\ &= e^{-\lambda t p}\frac{(\lambda t p)^n}{n!}\sum_{m=0}^{\infty} e^{-\lambda t(1-p)}\frac{(\lambda t(1-p))^m}{m!} \\ &= e^{-\lambda t p}\frac{(\lambda t p)^n}{n!}.\end{aligned}$$

In a similar fashion we compute the marginal probability $\mathbf{P}\{N_B(t)=m\}$, obtaining:

$$\mathbf{P}\{N_B(t)=m\} = e^{-\lambda t(1-p)}\frac{(\lambda t(1-p))^m}{m!}.$$

Hence, by (12.2) we have that

$$\mathbf{P}\{N_A(t)=n, N_B(t)=m\} = \mathbf{P}\{N_A(t)=m\}\cdot\mathbf{P}\{N_B(t)=m\}, \quad (12.3)$$

showing that the processes are independent. Now because the other conditions in the First Definition such as independent increments are also obviously satisfied, we have that $\{N_A(t), t \geq 0\}$ forms a Poisson process with rate λp and that $\{N_B(t), t \geq 0\}$ forms an independent Poisson process with rate $\lambda(1-p)$. ■

12.8 Uniformity

Theorem 12.15 *Given that one event of a Poisson process has occurred by time t, that event is equally likely to have occurred anywhere in $[0,t]$.*

Proof: Let T_1 denote the time of that one event:

$$\begin{aligned}\mathbf{P}\{T_1 < s \mid N(t)=1\} &= \frac{\mathbf{P}\{T_1 < s \text{ and } N(t)=1\}}{\mathbf{P}\{N(t)=1\}} \\ &= \frac{\mathbf{P}\{1 \text{ event in } [0,s] \text{ and } 0 \text{ events in } [s,t]\}}{\frac{e^{-\lambda t}(\lambda t)^1}{1!}} \\ &= \frac{\mathbf{P}\{1 \text{ event in } [0,s]\}\cdot\mathbf{P}\{0 \text{ events in } [s,t]\}}{e^{-\lambda t}\cdot\lambda t} \\ &= \frac{e^{-\lambda s}\cdot\lambda s\cdot e^{-\lambda(t-s)}\cdot(\lambda(t-s))^0}{e^{-\lambda t}\cdot\lambda t} \\ &= \frac{s}{t}.\end{aligned}$$

■

Generalization: If k events of a Poisson process occur by time t, then the k events are distributed independently and uniformly in $[0, t]$ [62, pp. 36–38].

12.9 Exercises

12.1 **Doubling Exponentials**
Suppose that job sizes are distributed $\text{Exp}(\mu)$. If job sizes all double, what can we say about the distribution of job sizes now? Prove it.

12.2 **Conditional Exponential**
Let $X \sim \text{Exp}(\lambda)$. What is $\mathbf{E}\left[X^2 \mid X < 1\right]$? [Hint: No integrals, just think!]

12.3 **Stationary and independent increments**
For a Poisson process with arrival rate λ, let $N(t)$ denote the number of arrivals by time t. Simplify the following, pointing out explicitly where you used stationary increments and where you used independent increments:

$$\mathbf{P}\{N(t) = 10 \mid N(3) = 2\} \qquad \text{(assume } t > 3\text{)}.$$

12.4 **Poisson process definition**
Suppose requests arrive to a website according to a Poisson process with rate $\lambda = 1$ request per ms. What is the probability that there are 5 arrivals in the first 5 ms and 10 arrivals in the first 10 ms?

12.5 **Packets of different colors**
(a) A stream of packets arrives according to a Poisson process with rate $\lambda = 50$ packets/s. Suppose each packet is of type "green" with probability 5% and of type "yellow" with probability 95%. Given that 100 green packets arrived during the previous second, (i) what is the expected number of yellow packets that arrived during the previous second? And (ii) what is the probability that 200 yellow packets arrived during the previous second?
(b) Red packets arrive according to a Poisson process with rate $\lambda_1 = 30$ packets/s. Black packets arrive according to a Poisson process with rate $\lambda_2 = 10$ packets/s. Assume the streams are merged into one stream. Suppose we are told that 60 packets arrived during one second. What is the probability that exactly 40 of those were red?

12.6 **Uniformity**
Packets arrive according to a Poisson process with rate λ. You are told that by time 30 seconds, 100 packets have arrived. What is the probability that 20 packets arrived during the first 10 seconds?

12.7 **Poisson process products**
Suppose customers arrive to a store according to a Poisson process with rate λ customers per second. Let $N(t)$ denote the number of arrivals by time t. What is $\mathbf{E}[N(s)N(t)]$, where $s < t$?

12.8 **Number of Poisson arrivals during S**
Let A_S denote the number of arrivals of a Poisson process with rate λ during S, where S is a continuous non-negative r.v., and the Poisson process is independent of S. You will derive $\mathbf{Var}(A_S)$ in two different ways:
(a) Do it without transforms.
(b) Derive the z-transform of A_S and differentiate it appropriately.

12.9 **Malware and honeypots**
A new malware is out in the Internet! We want to estimate its spread by time t. Internet hosts get infected by this malware according to a Poisson process with parameter λ, where λ is *not known*. Thrasyvoulos installs a honeypot security system to detect whether hosts are infected. Unfortunately there is a *lag time* between when a computer is infected and the honeypot detects the damage. Assume that this lag time is distributed $\text{Exp}(\mu)$. Suppose that the honeypot system has detected $N_1(t)$ infected hosts by time t. Thrasyvoulos worries that, because of the lag, the number of infected hosts is actually much higher than $N_1(t)$. We ask: How many *additional* hosts, $N_2(t)$, are expected to also be infected at time t.
(a) Suppose that an infection happens at time s, where $0 < s < t$. What is the probability that the infection is detected by time t?
(b) Consider an arbitrary infection that happens before time t. What is the (unconditional) probability, p, that the infection is detected by the honeypot by time t?
(c) How can we use our knowledge of $N_1(t)$ to estimate λ as a function of $N_1(t)$?
(d) Use your estimate of λ to determine the expected value of $N_2(t)$ as a function of $N_1(t)$.

12.10 **Sum of Geometric number of Exponentials**
Let $N \sim \text{Geometric}(p)$. Let $X_i \sim \text{Exp}(\mu)$. Let $S_N = \sum_{i=1}^{N} X_i$.
(a) What is the distribution of S_N? Prove this using a δ-step argument.
(b) Based on what you learned in (a), what is $\mathbf{P}\{S_N > t\}$?
(c) For a Poisson process with rate λ, where packets are colored "red" with probability q, what is the variance of the time between red packets?

12.11 **Reliability theory: max of two Exponentials**
Redundancy is often built into systems so that if a disk fails there is no catastrophe. The idea is to have the data on two disks, so that a catastrophe

only occurs if *both* disks fail. The time until a catastrophe occurs can be viewed as the "max" of two random variables.

(a) Let $X_1 \sim \text{Exp}(\lambda)$. Let $X_2 \sim \text{Exp}(\lambda)$. Suppose $X_1 \perp X_2$. What is $\mathbf{E}[\max(X_1, X_2)]$?

(b) Let $X_1 \sim \text{Exp}(\lambda_1)$. Let $X_2 \sim \text{Exp}(\lambda_2)$. Suppose $X_1 \perp X_2$. What is $\mathbf{E}[\max(X_1, X_2)]$?

12.12 **Exponential downloads**

You need to download two files: file 1 and file 2. File 1 is available via source A or source B. File 2 is available only via source C. The time to download file 1 from source A is Exp(1). The time to download file 1 from source B is Exp(2). The time to download file 2 from source C is Exp(3). You decide to download from *all three* sources simultaneously, in the hope that you get both file 1 and file 2 as soon as possible. Let T denote the time until you get *both* files.

(a) What is $\mathbf{E}[T]$?

(b) What is $\mathbf{P}\{T < t\}$?

12.13 **Reliability theory: max of many Exponentials**

Let $X_1, X_2, \ldots, X_n$ be i.i.d. with distribution $\text{Exp}(\lambda)$. Let

$$Z = \max(X_1, X_2, \ldots, X_n).$$

(a) What is $\mathbf{E}[Z]$?

(b) Roughly, what does $\mathbf{E}[Z]$ look like as a function of n and λ when n is reasonably high?

(c) Derive the distribution of Z.

12.14 **Conditional distribution**

Let $X \sim \text{Exp}(\lambda_X)$ and $Y \sim \text{Exp}(\lambda_Y)$, where $X \perp Y$. Let $Z = \min(X, Y)$. Prove that

$$(X \mid X < Y) \sim Z.$$

That is, show that $\mathbf{P}\{X > t \mid X < Y\} = \mathbf{P}\{Z > t\}$.

Before you start, take a minute to think about what this problem is saying: Suppose for simplicity that X and Y are both drawn from $\text{Exp}(\lambda)$. Say I put X in one hand and Y in the other, without looking. If you ask to see a random hand, the value you get is distributed $\text{Exp}(\lambda)$. However, if you ask me to look inside my hands and hand over the smaller of the two values, then the value that I give you will no longer be distributed $\text{Exp}(\lambda)$.

12.15 **Two two-stage jobs**

We have two jobs, X and Y, where each has two stages, as shown in Figure 12.6. Both stages of a job must be completed in order. That is, to complete job X, we need to first run X_1 and then run X_2. Similarly, to

complete job Y we must run Y_1 followed by Y_2. Assume that X_1, X_2, Y_1, and Y_2 are i.i.d. with distribution Exp(μ). Suppose that job X and job Y start running at the same time.

Figure 12.6 *Figure for Exercise 12.15.*

(a) What is the expected time until the first of these jobs completes?
(b) What is the expected time until the last of these jobs completes?

12.16 **Population modeling**

Naveen is interested in modeling population growth over time. He figures it is reasonable to model the birth process as a Poisson process with some average rate λ. He also assumes that a person's lifespan follows some distribution, T, with c.d.f. $F_T(t)$ and tail $\overline{F}_T(t) = 1 - F_T(t)$, where he assumes that lifespans of individuals are independent. Let $N(t)$ denote the population (number of people who are alive) at time t.

(a) Prove that $\mathbf{E}[N(t)] = \lambda \int_{k=0}^{t} \overline{F}_T(t-k)dk$.
(b) Naveen reads that approximately $\lambda = 4$ million people are born in the United States per year. He can't find a good distribution for lifespan, T, but he notes that the average life expectancy is $\mathbf{E}[T] = 75$ years. He decides to approximate lifespan by the Uniform(50, 100) distribution. Given these numbers, what can Naveen say about $\mathbf{E}[N(t)]$? Provide formulas for the three cases: $t < 50$; $50 < t < 100$; and $t > 100$.
(c) What does Naveen's model say about $\mathbf{E}[N(t)]$ as $t \to \infty$, meaning we're in steady state.