1 The Probability Distribution of the Age of Information

Yoshiaki Inoue, Tetsuya Takine, and Toshiyuki Tanaka

The freshness of information is the most important factor in designing real-time monitoring systems. The theory of the Age of Information (AoI) provides an explicit way to incorporate this perspective into the system design. This chapter is aimed at introducing the basic concept of the AoI and explaining its mathematical aspects. We first present a general introduction to the AoI and its standard analytical method. We then proceed to advanced material regarding the characterization of distributional properties of the AoI. Some bibliographical notes are also provided at the end of this chapter.

1.1 A General Introduction to the Age of Information

The Age of Information is a performance metric quantifying *the information freshness* in real-time monitoring systems. Let us consider a situation that a time-varying information source is monitored remotely (Figure 1.1). A sensor is attached to the information source and it observes (samples) the current status with some frequency. Each time the sensor samples the information source, it generates an update containing the obtained sample and sends it to a remote server, where some computational task is performed to extract state-information from raw data. The extracted information is then transmitted to a monitor and updates the information being displayed. The AoI A_t at time t is defined as *the elapsed time from the generation time* η_t *of the update whose information is displayed on the monitor at time* t:

$$A_t := t - \eta_t, \quad t \geq 0. \tag{1.1}$$

We thus have $\eta_t = t - A_t$, that is, the information displayed at time t was reported by an update generated at time $t - A_t$ by the sensor. In this sense, the AoI A_t directly quantifies the freshness of the information being displayed: the smaller value the AoI A_t takes, the fresher the information is. Because "the sampling of the information source" and "the generation of an update" occur simultaneously, these words could be used interchangeably. In the rest of this chapter, however, we shall describe the behavior of the system, consistently focusing only on the generation of updates.

By definition, η_t ($t \geq 0$) is a piecewise constant function of t: η_t jumps upward when the displayed information is updated, while it does not change its value elsewhere. Therefore, we see from (1.1) that A_t is a piecewise linear function of t with downward jumps at update instants. In particular, the AoI plotted along the time axis

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

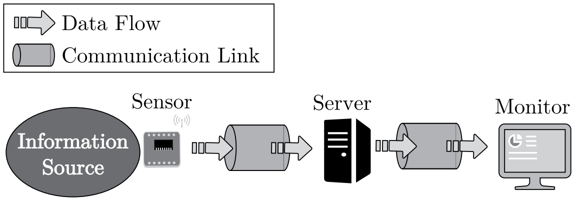


Figure 1.1 A remote monitoring system.

has a sawtooth graph as depicted in Figure 1.2, where $\beta_\ell^\dagger$ and $\alpha_\ell^\dagger$ ($\ell = 0, 1, \ldots$) represent the ℓth update time at the monitor and the generation time of that update at the sensor.

Note here that some updates generated by the sensor may *not* be displayed on the monitor forever because of the loss in communication links or subsequent updates' overtaking, where the latter is typically due to the management policy at the server (see Figure 1.1). Therefore, $(\alpha_\ell^\dagger)_{\ell=0,1,\ldots}$ and $(\beta_\ell^\dagger)_{\ell=0,1,\ldots}$ shown in Figure 1.2 are subsequences of $(\alpha_n)_{n=0,1,\ldots}$ of generation times of updates and $(\beta_n)_{n=0,1,\ldots}$ of their reception times, where $\beta_n = \infty$ if the update generated at time α_n is not displayed on the monitor forever. In this sense, we refer to $\alpha_\ell^\dagger$ and $\beta_\ell^\dagger$ as the generation and reception times of the ℓth *effective* update. A more formal discussion on overtaking of updates will be given in Section 1.3.1.

Let $G_\ell^\dagger$ and $D_\ell^\dagger$ denote the intergeneration time and the system delay of the ℓth effective update:

$$G_\ell^\dagger = \alpha_\ell^\dagger - \alpha_{\ell-1}^\dagger, \qquad D_\ell^\dagger = \beta_\ell^\dagger - \alpha_\ell^\dagger. \tag{1.2}$$

We observe from the definition of η_t that the AoI just after an update of the monitor equals the system delay experienced by the latest update:

$$A_{\beta_\ell^\dagger} = \beta_\ell^\dagger - \alpha_\ell^\dagger = D_\ell^\dagger.$$

On the other hand, the AoI just before an update of the monitor is called *the peak AoI*, as it corresponds to the peak of the sawtooth graph of the AoI process:

$$A_{\text{peak},\ell} := \lim_{t \to \beta_\ell^\dagger -} A_t = \beta_\ell^\dagger - \alpha_{\ell-1}^\dagger = D_\ell^\dagger + G_\ell^\dagger, \tag{1.3}$$

where the last equality follows from (1.2). We thus see that for each interval $[\beta_\ell^\dagger, \beta_{\ell+1}^\dagger)$ between the monitor's updates, the AoI process linearly increases from the system delay $D_\ell^\dagger$ to the peak AoI $A_{\text{peak},\ell+1}^\dagger$ with slope one (see Figure 1.2). This observation highlights the key difference between the AoI and the conventional delay metric: *The delay $D_\ell^\dagger$ represents only the information freshness immediately after an update* and it does not provide any information about the evolution of the information freshness between updates.

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

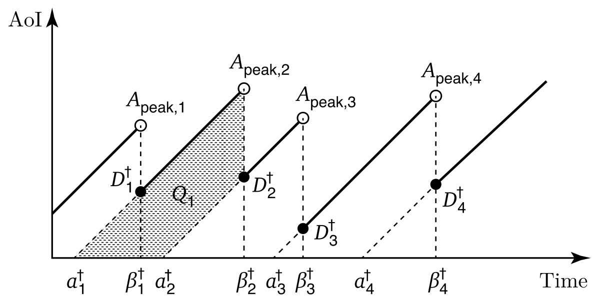


Figure 1.2 An example of the AoI process.

1.1.1 A Graphical Analysis of the Average AoI

Since the AoI $(A_t)_{t\geq 0}$ is a time-varying process, we need to consider some summary metric for the system performance, which is obtained by applying a functional to $(A_t)_{t\geq 0}$. The most commonly used summary metric is the time-averaged AoI defined as

$$m_A^\sharp := \lim_{T\to\infty} \frac{1}{T}\int_0^T A_t \mathrm{d}t. \tag{1.4}$$

The average AoI $m_A^\sharp$ can be analyzed with a graphical argument. Here, we provide only an informal argument to avoid technical complications; a more rigorous proof will be given in Section 1.3.

Observe that the area of the shaded trapezoid in Figure 1.2 is given by

$$Q_i = \frac{(A_{\text{peak},i+1})^2}{2} - \frac{(D_{i+1}^\dagger)^2}{2}.$$

By summing up Q_i for $i = 1, 2, \ldots$, we can calculate the area under the graph of the AoI, except for both ends. The boundary effects are negligible under suitable regularity conditions and we obtain

$$m_A^\sharp = \lim_{T\to\infty} \frac{M(T)}{T} \cdot \frac{1}{M(T)} \sum_{i=1}^{M(T)} Q_i = \lambda^\dagger \left\{ \frac{m_{(A_{\text{peak}})^2}}{2} - \frac{m_{(D^\dagger)^2}}{2} \right\}, \tag{1.5}$$

where $M(T) = \max\{n;\ \beta_n^\dagger \leq T\}$ denotes the total number of displayed information updates in time interval $(0, T]$, $\lambda^\dagger := \lim_{T\to\infty} M(T)/T$ denotes the average effective update rate, and

$$m_{(A_{\text{peak}})^2} := \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N (A_{\text{peak},i})^2, \quad m_{(D^\dagger)^2} := \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^N (D_i^\dagger)^2.$$

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

If the system is represented as a stationary and ergodic stochastic process,[1] (1.5) is equivalent to the following relation for the mean AoI and the second moments of the peak AoI and system delay:

$$\mathrm{E}[A] = \lambda^{\dagger} \cdot \frac{\mathrm{E}[(A_{\mathrm{peak}})^2] - \mathrm{E}[(D^{\dagger})^2]}{2}, \tag{1.6}$$

where A, A_{peak}, and $D^{\dagger}$ denote generic random variables for stationary A_t, $A_{\mathrm{peak},\ell}$, and $D_{\ell}^{\dagger}$. Furthermore, using (1.3) and noting the stationarity, (1.6) is rewritten as

$$\mathrm{E}[A] = \lambda^{\dagger} \left(\frac{\mathrm{E}[(G^{\dagger})^2]}{2} + \mathrm{E}\left[G_{\ell}^{\dagger} D_{\ell}^{\dagger} \right] \right), \tag{1.7}$$

where $G^{\dagger}$ denotes a generic random variable for stationary $G_{\ell}^{\dagger}$. It is worth noting that the peak AoI is also given in terms of the system delay $D_{\ell}^{\dagger}$ and the inter-update time $J_{\ell}^{\dagger} := \beta_{\ell}^{\dagger} - \beta_{\ell-1}^{\dagger}$ by

$$A_{\mathrm{peak},\ell+1} = D_{\ell}^{\dagger} + J_{\ell+1}^{\dagger},$$

so that we have yet another equivalent formula for $\mathrm{E}[A]$:

$$\mathrm{E}[A] = \lambda^{\dagger} \left(\frac{\mathrm{E}[(J^{\dagger})^2]}{2} + \mathrm{E}\left[D_{\ell}^{\dagger} J_{\ell+1}^{\dagger} \right] \right), \tag{1.8}$$

where $J^{\dagger}$ denotes a generic random variable for stationary $J_{\ell}^{\dagger}$. While these three formulas (1.6), (1.7), and (1.8) are equivalent, the choice of which expression to use in the analysis often affects the degree of tractability. In the following subsection, we briefly demonstrate applications of these formulas to first-come first-served (FCFS) single-server queueing models.

1.1.2 Queueing Modeling of Monitoring Systems

As we have seen, the AoI process $(A_t)_{t \geq 0}$ is characterized in terms of effective generation times $(\alpha_{\ell}^{\dagger})_{\ell=0,1,\ldots}$ of updates by the sensor and their reception times $(\beta_{\ell}^{\dagger})_{\ell=0,1,\ldots}$ by the monitor. Because $\beta_{\ell}^{\dagger} \geq \alpha_{\ell}^{\dagger}$ always holds by their definitions, we can think of *a queueing system* defined by the sequences of *arrival times* $(\alpha_{\ell}^{\dagger})_{\ell=0,1,\ldots}$ and *departure times* $(\beta_{\ell}^{\dagger})_{\ell=0,1,\ldots}$.[2] To be more specific, consider a virtual *service system*, where updates enter it immediately after their generations and leave it when they are received by the monitor (see Figure 1.3). This service system can be considered as an abstraction of a series of components that intermediate between the information source and the monitor. For example, it may represent a communication network to transfer update

[1] Stationarity refers to the property that probability distributions representing the system dynamics are time-invariant. Also, ergodicity refers to the property that time-averages coincide with corresponding ensemble averages. A more detailed explanation will be given in Section 1.3.3.

[2] In the later sections, we will consider a modeling that includes arrival and departure times of noneffective updates as well as effective updates.

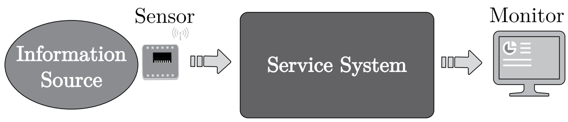


Figure 1.3 An abstraction of monitoring systems (cf. Figure 1.1).

packets, a server to extract status information from raw data, or a combination of them (see Figures 1.1 and 1.3).

The simplest model of the service system would be an FCFS single-server queue, that is, assuming that the service system consists of a first-in first-out (FIFO) buffer and a server. In what follows, we consider the mean AoI $\mathrm{E}[A]$ in two different FCFS single-server queues, where service times are assumed to follow an exponential distribution with mean $1/\mu$ ($\mu > 0$).

First, suppose that the intergeneration time $G^\dagger$ of effective updates is constant and equal to τ, that is, the D/M/1 queue in Kendall's notation. For the system stability, we assume $\tau > 1/\mu$. In this case, the formula in (1.7) is useful because we readily have $\lambda^\dagger = 1/\tau$, $\mathrm{E}[(G^\dagger)^2] = \tau^2$, and $\mathrm{E}[G^\dagger_\ell D^\dagger_\ell] = \tau \mathrm{E}[D^\dagger]$, yielding

$$\mathrm{E}[A] = \frac{\tau}{2} + \mathrm{E}[D^\dagger].$$

From the elementary queueing theory (Kleinrock 1975, p. 252), the system delay in the FCFS D/M/1 queue follows an exponential distribution with mean $1/\{\mu(1 - x^\star)\}$, where $x^\star$ is the unique solution of the following equation:

$$x = e^{-\tau\mu(1-x)}, \quad 0 < x < 1. \tag{1.9}$$

The mean AoI is thus given by

$$\mathrm{E}[A] = \frac{\tau}{2} + \frac{1}{\mu(1 - x^\star)}. \qquad \text{(D/M/1)}$$

Next, suppose that the sensor randomly generates effective updates according to a Poisson process with rate $\lambda^\dagger$, that is, the FCFS M/M/1 queue, where we assume $0 < \lambda^\dagger < \mu$ for stability. The intergeneration time $G^\dagger$ then follows an exponential distribution with mean $1/\lambda^\dagger$. In this case, the formula in (1.8) provides an easy way to calculate $\mathrm{E}[A]$. If an update does not depart the service system before the arrival of the next update (i.e., $D^\dagger_\ell \geq G^\dagger_{\ell+1}$), then the next service starts just after the departure, so that the next inter-departure time $J^\dagger_{\ell+1}$ equals a service time. If $D^\dagger_\ell < G^\dagger_{\ell+1}$, on the other hand, $J^\dagger_{\ell+1}$ equals the sum of the residual inter-arrival time (exponentially distributed with mean $1/\lambda^\dagger$ because of the memoryless property) and a service time. In both cases, $J^\dagger_{\ell+1}$ is conditionally independent of $D^\dagger_\ell$, given either $D^\dagger_\ell \geq G^\dagger_{\ell+1}$ or $D^\dagger_\ell < G^\dagger_{\ell+1}$. Using the fact that the system delay in the FCFS M/M/1 queue follows an exponential distribution with mean $1/(\mu - \lambda^\dagger)$ (Kleinrock 1975, p. 205), we obtain

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

$$\begin{aligned}
\mathrm{E}\left[D_\ell^\dagger J_{\ell+1}^\dagger\right] &= \mathrm{E}\left[\mathbb{1}\{D_\ell^\dagger \geq G_{\ell+1}^\dagger\}D_\ell^\dagger J_{\ell+1}^\dagger\right] + \mathrm{E}\left[\mathbb{1}\{D_\ell^\dagger < G_{\ell+1}^\dagger\}D_\ell^\dagger J_{\ell+1}^\dagger\right] \\
&= \int_0^\infty x \cdot \frac{1}{\mu} \cdot (1 - e^{-\lambda^\dagger x}) \cdot (\mu - \lambda^\dagger)e^{-(\mu-\lambda^\dagger)x}\mathrm{d}x \\
&\qquad + \int_0^\infty x \cdot \left(\frac{1}{\lambda^\dagger} + \frac{1}{\mu}\right) \cdot e^{-\lambda^\dagger x} \cdot (\mu - \lambda^\dagger)e^{-(\mu-\lambda^\dagger)x}\mathrm{d}x \\
&= \frac{1}{\mu\lambda^\dagger}\left(1 - \rho + \frac{\rho}{1-\rho}\right),
\end{aligned} \tag{1.10}$$

and

$$\begin{aligned}
\mathrm{E}\left[(J^\dagger)^2\right] &= \mathrm{E}\left[\mathbb{1}\{D_\ell^\dagger \geq G_{\ell+1}^\dagger\}(J^\dagger)^2\right] + \mathrm{E}\left[\mathbb{1}\{D_\ell^\dagger < G_{\ell+1}^\dagger\}(J^\dagger)^2\right] \\
&= \int_0^\infty \frac{2}{\mu^2} \cdot (1 - e^{-\lambda^\dagger x}) \cdot (\mu - \lambda^\dagger)e^{-(\mu-\lambda^\dagger)x}\mathrm{d}x \\
&\qquad + \int_0^\infty \left(\frac{2}{(\lambda^\dagger)^2} + \frac{2}{\mu\lambda^\dagger} + \frac{2}{\mu^2}\right) \cdot e^{-\lambda^\dagger x}(\mu - \lambda^\dagger)e^{-(\mu-\lambda^\dagger)x}\mathrm{d}x \\
&= \frac{2\rho^2}{\mu\lambda^\dagger} + \frac{2(1-\rho)}{(\lambda^\dagger)^2} \cdot (1 + \rho + \rho^2),
\end{aligned} \tag{1.11}$$

where $\rho := \lambda^\dagger/\mu$ denotes the traffic intensity. Therefore, we obtain from (1.8), (1.10), and (1.11),

$$\mathrm{E}[A] = \frac{1}{\mu}\left(1 + \frac{1}{\rho} + \frac{\rho^2}{1-\rho}\right). \qquad \text{(M/M/1)}$$

We conclude this section by presenting some numerical examples. We set the time unit so that $\mu = 1$ holds throughout. Figure 1.4 shows the mean AoI $\mathrm{E}[A]$ as a function of the generation rate $\lambda^\dagger$ (i.e., the rate at which effective updates are generated). We observe that $\mathrm{E}[A]$ forms a U-shaped curve with respect to $\lambda^\dagger$. This U-shaped curve of $\mathrm{E}[A]$ is understood to be due to *the trade-off between the generation interval $G^\dagger$ and the system delay $D^\dagger$*: while reducing the generation interval would be effective in keeping the information fresher, it would also increase the delay in the service system. To illustrate this fact, in Figure 1.4, we also plot the mean backward recurrence time $\mathrm{E}[(G^\dagger)^2]/(2\mathrm{E}[G^\dagger])$ of the generation process (i.e., the expected elapsed time since the last generation instant) and the mean system delay $\mathrm{E}[D]$. Note here that the sum of these terms $\mathrm{E}[(G^\dagger)^2]/(2\mathrm{E}[G^\dagger]) + \mathrm{E}[D]$ can be regarded as an approximation to the mean AoI $\mathrm{E}[A]$ ignoring the dependence between $G_\ell^\dagger$ and $D_\ell^\dagger$ (cf. (1.7)). We observe that the former term is dominant for small $\lambda^\dagger$, whereas the latter term becomes dominant for large values of $\lambda^\dagger$. The mean AoI $\mathrm{E}[A]$ is then minimized at a moderate value of $\lambda^\dagger$, where these effects on $\mathrm{E}[A]$ are appropriately balanced.

Figure 1.5 compares the mean AoI in the FCFS M/M/1 and D/M/1 queues. We observe that constant generation intervals are preferable to exponential generation intervals in terms of the mean AoI. Intuitively, the superiority of constant generation intervals can be understood as follows. Firstly, the constant generation intervals

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

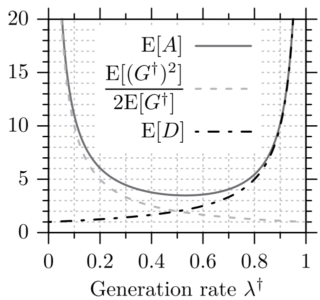


Figure 1.4 The mean AoI E[A] in the FCFS M/M/1 queue.

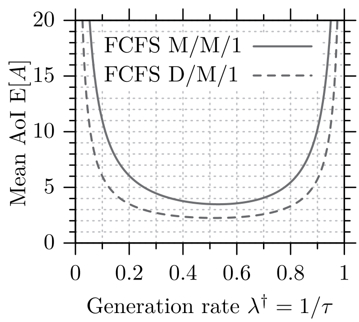


Figure 1.5 The mean AoI E[A] in the FCFS M/M/1 and D/M/1 queues.

minimize the mean backward recurrence time of the generation process, which can be verified with

$$\frac{\mathrm{E}[(G^\dagger)^2]}{2\mathrm{E}[G^\dagger]} = \frac{\mathrm{Var}[G^\dagger] + (\mathrm{E}[G^\dagger])^2}{2\mathrm{E}[G^\dagger]} = \frac{1 + (\mathrm{Cv}[G^\dagger])^2}{2} \cdot \mathrm{E}[G^\dagger],$$

where $\mathrm{Cv}[Y] = \sqrt{\mathrm{Var}[Y]}/\mathrm{E}[Y]$ denotes the coefficient of variation of random variable Y. Because we have $\mathrm{Cv}[G^\dagger] \geq 0$ and the equality holds if and only if $G^\dagger$ is constant, the mean backward recurrence time is minimized by constant generation intervals. Secondly, the mean delay $1/\{\mu(1 - x^\star)\}$ in the FCFS D/M/1 queue is smaller than the mean delay $1/\{\mu(1 - \rho)\}$ in the FCFS M/M/1 queue because $x^\star$ defined by (1.9) is smaller than ρ.

1.2 The Probability Distribution of the AoI

The rest of this chapter is devoted to discussions on the *probability distribution* of the AoI. As mentioned in the previous section, we need to use some summary metric of the AoI process $(A_t)_{t\geq 0}$ for performance evaluation. The *asymptotic frequency distribution (AFD)* of the AoI process is considered to be one of the most fundamental

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

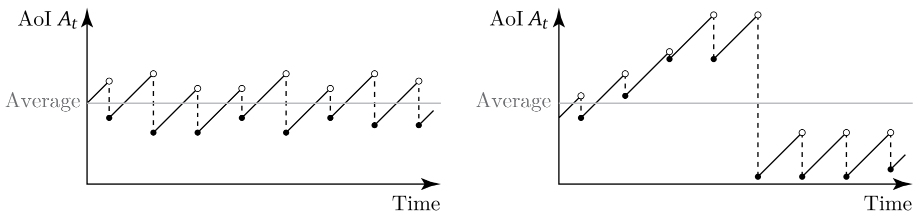


Figure 1.6 An example of different AoI processes with equal time-averages.

quantities among various kinds of summary metrics. The AFD $F_A^\sharp(x)$ of the AoI process is defined as

$$F_A^\sharp(x) = \lim_{T\to\infty} \frac{1}{T}\int_0^T \mathbb{1}\{A_t \le x\}\mathrm{d}t, \quad x \ge 0,$$

where $\mathbb{1}\{\cdot\}$ denotes an indicator function. The value of the AFD for fixed $x \ge 0$ thus represents the long-run fraction of time that the AoI does not exceed the threshold x. As in the previous section, the system is usually modeled as a stationary and ergodic stochastic process in theoretical studies on the AoI. Within such a framework, the AFD can be equated with the probability distribution of the stationary AoI A, which has the same distribution as A_t for all $t \ge 0$:

$$F_A^\sharp(x) = \Pr(A \le x) = \Pr(A_t \le x).$$

For the time being, we again focus on the stationary and ergodic system. We will provide a more detailed discussion on this point later in Section 1.3.

The probability distribution of the AoI (the AoI distribution in short) has several appealing properties in characterizing AoI performances. Firstly, although the time-averaged AoI (1.4) is the most widely used summary metric, it has a serious weakness as a performance metric: It cannot capture *how the information freshness fluctuates over time*. Figure 1.6 depicts an example of two AoI processes, which differ substantially in their fluctuations, but cannot be distinguished by the time-average alone. On the other hand, the AoI distribution contains much information about the fluctuation of the process. For example, a standard method to quantify the degree of fluctuation is to use the variance of the process

$$(\sigma_A^\sharp)^2 := \lim_{T\to\infty} \frac{1}{T}\int_0^T \left(A_t - m_A^\sharp\right)^2 \mathrm{d}t,$$

which can be readily computed from the probability distribution:

$$m_A^\sharp = \int_0^\infty x\mathrm{d}F_A(x),$$
$$(\sigma_A^\sharp)^2 = \int_0^\infty \left(x - m_A^\sharp\right)^2 \mathrm{d}F_A(x) = \int_0^\infty x^2\mathrm{d}F_A(x) - \left(\int_0^\infty x\mathrm{d}F_A(x)\right)^2.$$

Also, another common way to capture the variability is the use of a box-and-whisker diagram, which will be demonstrated in Section 1.4.

Secondly, it is often the case that the cost of stale information increases *nonlinearly* as time passes since its generation. The AoI value averaged with a nonlinear cost function $g(\cdot)$ is of interest in such a situation:

$$m^{\sharp}_{g(A)} := \lim_{T\to\infty} \frac{1}{T} \int_0^T g(A_t)\mathrm{d}t. \tag{1.12}$$

The AoI distribution $F_A(\cdot)$ provides a simple way to calculate the average nonlinear cost:

$$m^{\sharp}_{g(A)} = \int_0^{\infty} g(x)\mathrm{d}F_A(x),$$

that is, an analysis of the AoI process with any nonlinear cost functions reduces to that of the AoI distribution. An indicator function $g(y) = \mathbb{1}\{y > \theta\}$ with a threshold θ is a particular example of a nonlinear cost function, whose time-average is equal to the value of the complementary AoI distribution $\overline{F}_A(x) := 1 - F_A(x)$ evaluated at $x = \theta$. This form of the cost function is of interest for system reliability because making $\overline{F}_A(\theta) < \epsilon$ be satisfied for small ϵ guarantees that the AoI value is below the threshold θ for a fraction $(1 - \epsilon)$ of the time.

Thirdly, the AoI distribution is useful in characterizing monitoring errors when the dynamics of the information source is specified. Suppose that the information source is represented as a stochastic process $(X_t)_{t\geq 0}$ and that the monitor displays the latest state information $\hat{X}_t$ received:

$$\hat{X}_t = X_{t-A_t}.$$

Assuming that the AoI $(A_t)_{t\geq 0}$ and the monitored state $(X_t)_{t\geq 0}$ are independent,[3] the expected error measured with a penalty function $L(\cdot,\cdot)$ is given by

$$\mathrm{E}\left[L(X_t, \hat{X}_t)\right] = \mathrm{E}\left[L(X_t, X_{t-A_t})\right] = \int_0^{\infty} \mathrm{E}\left[L(X_t, X_{t-x})\right] \mathrm{d}F_A(x),$$

that is, it is represented in terms of the AoI distribution $F_A(\cdot)$, given the knowledge about the transition dynamics of X_t. This quantity further equals the time-average of the error $L(X_t, \hat{X}_t)$ if the monitored process $(X_t)_{t\geq 0}$ is also ergodic.

Finally, the AoI distribution will play an important role in developing statistical theory of the AoI, which would allow us to perform such tasks as parameter estimation, hypothesis testing, model selection, and so on the basis of a collection of observed AoI data, under situations where one does not have complete knowledge about the AoI-generating process of interest. The usefulness of the AoI distribution in this context as well is ascribed to the fact that a distribution is far more informative than a small number of statistics (summary metrics).

Since one cannot generally expect to explicitly write down the likelihood function for common models of the AoI process discussed in the literature, one would perform parameter estimation by first taking a set of statistics, evaluating their values on the basis of the observed AoI data, and then estimating the model parameters therefrom.

[3] This is usually the case if generation timings of updates are determined independently of the state X_t.

As an AoI model defines a mapping from its parameter space to the space of the statistics, parameter estimation amounts to evaluating the inverse of this mapping. It is, however, only possible if the forward mapping is one-to-one. Several existing AoI models, on the other hand, have more than one parameter, so that any single statistic, like the mean AoI, should be insufficient for parameter estimation, and one would therefore require at least as many statistics as the number of parameters in the model. Knowledge of the AoI distribution not only allows us to evaluate the forward mapping for a selected set of statistics, but furthermore would provide us with guidance on how to choose the statistics to be used in parameter estimation.

In model selection, in its simplest form, one takes two alternative AoI models, and decides which of the two models better explains the observed AoI data. For successful model selection, it is desirable that the ranges of the forward mappings associated with the two models are disjoint and well-separated. Such properties also depend on the choice of the statistics to be used, and knowledge of the AoI distributions plays an essential role here as well.

It should be noted that for a full-fledged statistical theory of the AoI one should go beyond the AoI distribution: in statistical procedures, such as obtaining confidence interval, performing hypothesis testing, and so on, one usually requires knowledge about distributions of the statistics evaluated on the basis of a finite-sized dataset from a prescribed AoI model. For example, in order to decide how reliable an estimate of the mean AoI from a finite dataset is, one would need to evaluate its variance, which in turn requires knowledge of the autocorrelation of the AoI process (Bhat & Rao 1987). In this regard, a full statistical theory of AoI is yet to be explored, and the AoI distribution may be recognized as a first step toward this direction.

1.3 A General Formula for the AoI Distribution

1.3.1 Model Description

We start by providing a formal description of the mathematical model to be considered. We suppose that updates are generated by the sensor with generation intervals $(G_n)_{n=1,2,\ldots}$. The sequence of generation times $(\alpha_n)_{n=0,1,\ldots}$ is then determined by the initial generation time α_0 and the recursion

$$\alpha_{n+1} = \alpha_n + G_{n+1}, \quad n = 0, 1, \ldots.$$

We refer to the update generated at time α_n as the nth update. Just after its generation, the nth update arrives at the service system (cf. Figure 1.3), which imposes a delay of length D_n. The information contained in the nth update is thus received by the monitor at time β_n, where

$$\beta_n = \alpha_n + D_n, \quad n = 0, 1, \ldots.$$

Without loss of generality, we set the time origin so that $\beta_0 \leq 0$.

The AoI process $(A_t)_{t\geq 0}$ is then constructed as follows. We first note that the AoI refers to the elapsed time of the *latest* status information displayed on the monitor.

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

That is, the AoI process is not affected by status updates overtaken by newer updates. More formally, let $\mathcal{I}$ denote the index set of updates that are not overtaken by other updates:

$$\mathcal{I} = \{n;\ \beta_n < \min\{\beta_{n+1}, \beta_{n+2}, \ldots\}\}. \tag{1.13}$$

Also, let $\mathcal{I}^{\mathrm{c}}$ denote the complement of $\mathcal{I}$:

$$\mathcal{I}^{\mathrm{c}} = \{0, 1, \ldots\} \setminus \mathcal{I}.$$

An update in $\mathcal{I}$ (resp. $\mathcal{I}^{\mathrm{c}}$) is said to be *effective* (resp. *noneffective*) in the sense that its information is newer (resp. older) than that displayed on the monitor just before its reception by the monitor update. As shown in what follows, the AoI process is completely characterized in terms of the generation times $(\alpha_n)_{n\in\mathcal{I}}$ and the reception times $(\beta_n)_{n\in\mathcal{I}}$ of effective updates only.

Recall that the AoI at time t is given by (1.1) in terms of η_t, which denotes the generation time of the latest information displayed on the monitor at time t. In the current setting, η_t is written as

$$\eta_t = \sup\{\alpha_n\,;\ n \in \{0, 1, \ldots\},\ \beta_n \le t\}.$$

For each noneffective update $i \in \mathcal{I}^{\mathrm{c}}$, there exists an integer $k \ge i+1$ such that $\beta_k \le \beta_i$ (cf. Eq. (1.13)), that is, $\beta_i \le t$ implies the existence of an update k with $\beta_k \le t$ and $\alpha_k \ge \alpha_i$. Therefore, the value of η_t is not affected by excluding all noneffective updates from consideration:

$$\eta_t = \sup\{\alpha_n\,;\ n \in \mathcal{I},\ \beta_n \le t\}. \tag{1.14}$$

We thus restrict our attention to the effective updates only. Recall that we use the superscript "†" to represent quantities of effective updates. Let $(\alpha_\ell^\dagger)_{\ell=0,1,\ldots}$ and $(\beta_\ell^\dagger)_{\ell=0,1,\ldots}$ denote the sequences of effective generation and reception times:

$$(\alpha_\ell^\dagger)_{\ell=0,1,\ldots} = (\alpha_n)_{n\in\mathcal{I}}, \quad (\beta_\ell^\dagger)_{\ell=0,1,\ldots} = (\beta_n)_{n\in\mathcal{I}}.$$

Also, we define the intergeneration time $G_\ell^\dagger$ ($\ell = 1, 2, \ldots$) and the system delay $D_\ell^\dagger$ ($\ell = 0, 1, \ldots$) of the ℓth effective update as in (1.2). Note here that while the effective system delay $D_\ell^\dagger$ equals the original system delay D_n for some $n \in \mathcal{I}$, the effective intergeneration time $G_\ell^\dagger$ is given by the sum of intergeneration times

$$G_\ell^\dagger = G_{n+1} + G_{n+2} + \cdots + G_{n+k+1}, \tag{1.15}$$

for some $n \in \mathcal{I}$ such that $n+1 \in \mathcal{I}^{\mathrm{c}}, n+2 \in \mathcal{I}^{\mathrm{c}}, \ldots, n+k \in \mathcal{I}^{\mathrm{c}}$, and $n+k+1 \in \mathcal{I}$.

From the construction of the sequence of effective updates, it is clear that

$$\ell < \ell' \;\Rightarrow\; \beta_\ell^\dagger < \beta_{\ell'}^\dagger,$$

that is, the effective updates enter and depart the service system in a first-in first-out (FIFO) manner. In other words, during the time interval $[\beta_\ell^\dagger, \beta_{\ell+1}^\dagger)$ between the consecutive receptions of the ℓth and $(\ell+1)$st effective updates, the ℓth update's information is the latest at the monitor. With this observation, (1.14) is considerably simplified as

$$\eta_t = \alpha_\ell^\dagger, \qquad t \in [\beta_\ell^\dagger, \beta_{\ell+1}^\dagger).$$

Therefore, the expression (1.1) for the AoI is rewritten as follows:

$$A_t = t - \alpha_\ell^\dagger, \qquad t \in [\beta_\ell^\dagger, \beta_{\ell+1}^\dagger),\ \ell = 0, 1, \ldots. \tag{1.16}$$

Our assumption $\beta_0 \le 0$ implies $\beta_0^\dagger \le 0$, so that the AoI A_t is well-defined for all $t \in [0, \beta_\infty^\dagger)$, where $\beta_\infty^\dagger := \lim_{\ell\to\infty} \beta_\ell^\dagger$. Note that we must have $\beta_\infty^\dagger = \infty$ in practical situations because otherwise there exists $T_{\mathrm{sup}} < \infty$ such that the monitor will never be updated again after time $t = T_{\mathrm{sup}}$.

Finally, recall that the ℓth peak AoI $A_{\mathrm{peak},\ell}$ is defined as the AoI just before the reception of the ℓth update, and it is given by (1.3).

1.3.2 The Asymptotic Frequency Distribution (AFD) of the AoI

In this subsection, we present a sample-path analysis of the AoI process. Mathematical analysis in this subsection deals with a *deterministic* (i.e., not random) sequence of generation intervals $(G_n)_{n=1,2,\ldots}$, system delays $(D_n)_{n=0,1,\ldots}$, and a deterministic value of the initial generation time α_0. As we have seen in the previous subsection, these quantities completely determine the AoI process $(A_t)_{t\ge 0}$, the peak AoI process $(A_{\mathrm{peak},\ell})_{\ell=1,2,\ldots}$, and the effective system delay process $(D_\ell^\dagger)_{\ell=0,1,\ldots}$. In the next subsection, we will turn our attention to a stochastic version of the AoI process, which is usually dealt with in the AoI literature.

The main purpose of this subsection is to derive a general relation satisfied by the AFDs of the (deterministic) AoI process $(A_t)_{t\ge 0}$, peak AoI process $(A_{\mathrm{peak},\ell})_{\ell=1,2,\ldots}$, and effective system delay process $(D_\ell^\dagger)_{\ell=0,1,\ldots}$. As mentioned previously, the AFD of a process is defined for each $x \ge 0$ as the long-run fraction of time that the process does not exceed x. More specifically, the AFDs of the AoI, peak AoI, and the effective system delay are defined respectively as

$$F_A^\sharp(x) = \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbb{1}\{A_t \le x\}\mathrm{d}t, \qquad x \ge 0, \tag{1.17}$$

$$F_{A_{\mathrm{peak}}}^\sharp(x) = \lim_{N\to\infty} \frac{1}{N} \sum_{\ell=1}^{N} \mathbb{1}\{A_{\mathrm{peak},\ell} \le x\}, \qquad x \ge 0, \tag{1.18}$$

$$F_{D^\dagger}^\sharp(x) = \lim_{N\to\infty} \frac{1}{N} \sum_{\ell=0}^{N-1} \mathbb{1}\{D_\ell^\dagger \le x\}, \qquad x \ge 0, \tag{1.19}$$

provided that these limits exist.

In order to obtain nontrivial results from the analysis, we need to impose several basic assumptions: (i) finiteness and positivity of the effective generation rate, (ii) rate stability of the FIFO sequence of effective updates, and (iii) the existence of the AFDs of the peak AoI and system delay. To be more specific, let $\lambda^\dagger$ denote the mean effective generation rate:

$$\lambda^\dagger := \lim_{T\to\infty} \frac{1}{T} \sum_{\ell=1}^{\infty} \mathbb{1}\{\alpha_\ell^\dagger \le T\} = \lim_{N\to\infty} \left(\frac{1}{N} \sum_{\ell=1}^{N} G_\ell^\dagger \right)^{-1}, \tag{1.20}$$

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

provided the limits exist. We note that the second equality of this equation implies that the mean effective generation rate equals the reciprocal of the mean effective generation interval of updates; a formal proof of this intuitive relation can be given with a deterministic version of the elementary renewal theorem (El-Taha & Stidham Jr. 1999, Lemma 1.1).

The assumptions just mentioned are then formally stated as follows:

ASSUMPTION 1.1

(i) The effective generation rate satisfies $\lambda^{\dagger} \in (0, \infty)$.
(ii) The effective update rate equals the effective generation rate:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{\ell=1}^{\infty} \mathbb{1}\{\beta_\ell^{\dagger} \le T\} = \lambda^{\dagger}. \tag{1.21}$$

(iii) The limits in (1.18) and (1.19) exist.

The key observation in our analysis is that for each interval $[\beta_\ell^{\dagger}, \beta_{\ell+1}^{\dagger})$ between effective updates at the monitor, the contribution of the AoI process to its frequency distribution (for a fixed $x \ge 0$) is represented as

$$\begin{aligned}\int_{\beta_\ell^{\dagger}}^{\beta_{\ell+1}^{\dagger}} \mathbb{1}\{A_t \le x\}\mathrm{d}t &= \int_{\beta_\ell^{\dagger}}^{\beta_{\ell+1}^{\dagger}} \mathbb{1}\{t - \alpha_\ell^{\dagger} \le x\}\mathrm{d}t \\ &= \int_{\beta_\ell^{\dagger}-\alpha_\ell^{\dagger}}^{\beta_{\ell+1}^{\dagger}-\alpha_\ell^{\dagger}} \mathbb{1}\{u \le x\}\mathrm{d}u \\ &= \int_{D_\ell^{\dagger}}^{A_{\text{peak},\ell+1}} \mathbb{1}\{u \le x\}\mathrm{d}u \\ &= \int_0^x \mathbb{1}\{A_{\text{peak},\ell+1} > u\}\mathrm{d}u - \int_0^x \mathbb{1}\{D_\ell^{\dagger} > u\}\mathrm{d}u, \end{aligned} \tag{1.22}$$

where we have the first equality from (1.16), the second equality by letting $u = t - \alpha_\ell^{\dagger}$, the third equality from (1.2) and (1.3), and the last equality from the following identity:

$$\int_0^y \mathbb{1}\{u \le x\}\mathrm{d}u = \int_0^{\infty} \mathbb{1}\{u \le x\}\mathbb{1}\{u < y\}\mathrm{d}u = \int_0^x \mathbb{1}\{y > u\}\mathrm{d}u, \quad x \ge 0,\ y \ge 0.$$

Using $\mathbb{1}\{y > x\} = 1 - \mathbb{1}\{y \le x\}$, we further rewrite (1.22) as

$$\int_{\beta_\ell^{\dagger}}^{\beta_{\ell+1}^{\dagger}} \mathbb{1}\{A_t \le x\}\mathrm{d}t = \int_0^x \mathbb{1}\{D_\ell^{\dagger} \le u\}\mathrm{d}u - \int_0^x \mathbb{1}\{A_{\text{peak},\ell+1} \le u\}\mathrm{d}u.$$

Therefore, we obtain the following relation by summing up both sides of the above equation for $\ell = 0, 1, \ldots, N-1$:

$$\int_{\beta_0^{\dagger}}^{\beta_N^{\dagger}} \mathbb{1}\{A_t \le x\}\mathrm{d}t = \int_0^x \sum_{\ell=0}^{N-1} \mathbb{1}\{D_\ell^{\dagger} \le u\}\mathrm{d}u - \int_0^x \sum_{\ell=1}^{N} \mathbb{1}\{A_{\text{peak},\ell} \le u\}\mathrm{d}u,$$

which is equivalent to that for $T > 0$,

$$\frac{1}{T}\int_{\beta_0^\dagger}^{\beta_N^\dagger} \mathbb{1}\{A_t \le x\}\mathrm{d}t$$
$$= \frac{N}{T}\left\{\int_0^x \frac{1}{N}\sum_{\ell=0}^{N-1} \mathbb{1}\{D_\ell^\dagger \le u\}\mathrm{d}u - \int_0^x \frac{1}{N}\sum_{\ell=1}^{N} \mathbb{1}\{A_{\mathrm{peak},\ell} \le u\}\mathrm{d}u\right\}. \qquad (1.23)$$

Letting $T = \beta_N^\dagger - \beta_0^\dagger$, we then see that this equation relates the frequency distribution of the AoI to those of the effective system delay and the peak AoI for the finite time interval $[\beta_0^\dagger, \beta_N^\dagger)$, using the effective update rate N/T in that interval.

With Assumption 1.1, this relation is further extended to its limiting version, which is the main result of this subsection:

LEMMA 1.2 *Under Assumption 1.1, the AFD of the AoI $F_A^\sharp(x)$ is related to those of the peak AoI $F_{A_{\mathrm{peak}}}^\sharp(x)$ and effective system delay $F_{D^\dagger}^\sharp(x)$ as*

$$F_A^\sharp(x) = \lambda^\dagger \int_0^x \left(F_{D^\dagger}^\sharp(y) - F_{A_{\mathrm{peak}}}^\sharp(y)\right)\mathrm{d}y, \quad x \ge 0,$$

where $\lambda^\dagger$ denotes the mean generation rate defined as in (1.20).

Proof Let $M(t)$ denote the total number of effective updates at the monitor in a time interval $(\beta_0^\dagger, t]$:

$$M(t) = \sup\{\ell \in \{0,1,\ldots\};\ \beta_\ell^\dagger \le t\} = \sum_{\ell=1}^{\infty} \mathbb{1}\{\beta_\ell^\dagger \le T\}.$$

For an interval $[0, T)$, the frequency distribution of the AoI is then written as

$$\frac{1}{T}\int_0^T \mathbb{1}\{A_t \le x\}\mathrm{d}t$$
$$= \frac{1}{T}\int_{\beta_0^\dagger}^{\beta_{M(T)}^\dagger} \mathbb{1}\{A_t \le x\}\mathrm{d}t - \frac{1}{T}\int_{\beta_0^\dagger}^{0} \mathbb{1}\{A_t \le x\}\mathrm{d}t + \frac{1}{T}\int_{\beta_{M(T)}^\dagger}^{T} \mathbb{1}\{A_t \le x\}\mathrm{d}t$$
$$= \frac{M(T)}{T}\cdot S_T(x) + \frac{\epsilon_T(x)}{T},$$

where $S_T(x)$ and $\epsilon_T(x)$ are defined as follows (cf. (1.23)):

$$S_T(x) = \int_0^x \frac{1}{M(T)}\sum_{\ell=0}^{M(T)-1} \mathbb{1}\{D_\ell^\dagger \le u\}\mathrm{d}u - \int_0^x \frac{1}{M(T)}\sum_{\ell=1}^{M(T)} \mathbb{1}\{A_{\mathrm{peak},\ell} \le u\}\mathrm{d}u, \quad (1.24)$$

$$\epsilon_T(x) = -\int_{\beta_0^\dagger}^{0} \mathbb{1}\{A_t \le x\}\mathrm{d}t + \int_{\beta_{M(T)}^\dagger}^{T} \mathbb{1}\{A_t \le x\}\mathrm{d}t.$$

To prove Lemma 1.2, it is then sufficient to show that

$$\lim_{T\to\infty} \frac{M(T)}{T} = \lambda^{\dagger}, \tag{1.25}$$

$$\lim_{T\to\infty} S_T(x) = \int_0^x \big(F^{\sharp}_{D^{\dagger}}(y) - F^{\sharp}_{A_{\text{peak}}}(y)\big)\mathrm{d}y, \tag{1.26}$$

$$\lim_{T\to\infty} \frac{\epsilon_T(x)}{T} = 0. \tag{1.27}$$

Note first that (1.25) immediately follows from Assumption 1.1 (ii). This further implies $M(T) \to \infty$ as $T \to \infty$, since we have $\lambda^{\dagger} > 0$ from Assumption 1.1 (i). We thus readily obtain (1.26) from (1.18), (1.19), and (1.24) using the dominated convergence theorem.

We then consider $\epsilon_T(x)$. By definition of $\epsilon_T(x)$ and $M(T)$, we have

$$\left|\frac{\epsilon_T(x)}{T}\right| \le \left|\frac{\beta_0^{\dagger}}{T}\right| + \left|\frac{\beta_{M(T)+1} - \beta_{M(T)}^{\dagger}}{T}\right|.$$

The first term on the right-hand side converges to zero as $T \to \infty$ because $|\beta_0^{\dagger}| < \infty$. It thus suffices to consider the second term. Similarly to (1.20), we have from the deterministic version of the renewal theorem (El-Taha & Stidham Jr. 1999, Lemma 1.1)

$$\begin{aligned}\left(\lim_{T\to\infty} \frac{M(T)}{T}\right)^{-1} &= \lim_{N\to\infty} \frac{1}{N}\sum_{\ell=1}^{N}(\beta_{\ell}^{\dagger} - \beta_{\ell-1}^{\dagger}) \\ &= \lim_{T\to\infty} \frac{1}{M(T)}(\beta_{M(T)}^{\dagger} - \beta_1^{\dagger}) \\ &= \lim_{T\to\infty} \frac{T}{M(T)} \cdot \frac{\beta_{M(T)}^{\dagger}}{T},\end{aligned}$$

which together with (1.25) imply

$$\lim_{T\to\infty} \frac{\beta_{M(T)}^{\dagger}}{T} = 1.$$

We then have

$$\lim_{T\to\infty} \left|\frac{\beta_{M(T)+1} - \beta_{M(T)}^{\dagger}}{T}\right| = 0,$$

so that (1.27) holds. □

1.3.3 The Stationary AoI Distribution in Ergodic Systems

In the previous subsection, we have derived Lemma 1.2 assuming deterministic sequences of intergeneration times $(G_n)_{n=1,2,\ldots}$ and system delays $(D_n)_{0,1,\ldots}$, and a deterministic value of the initial generation time α_0. In this subsection, we investigate

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

implications of this result to status update systems formulated as ergodic stochastic models. In the rest of this chapter, we follow a convention that for any nonnegative random variable Y, the cumulative distribution function (CDF) is denoted by $F_Y(x)$ ($x \geq 0$), the probability density function (if it exists) is denoted by $f_Y(x)$ ($x \geq 0$), and the Laplace–Stieltjes transform (LST) is denoted by $f_Y^*(s)$ ($s > 0$):

$$F_Y(x) := \Pr(Y \leq x), \quad f_Y(x) = \frac{\mathrm{d}F_Y(x)}{\mathrm{d}x}, \quad f_Y^*(s) = \mathrm{E}[e^{-sY}] = \int_0^\infty e^{-sx}\mathrm{d}F_Y(x).$$

Suppose that intergeneration times $(G_n)_{n=1,2,\ldots}$, system delays $(D_n)_{0,1,\ldots}$, and the initial generation time α_0 are given as random variables defined on a common probability space $(\Omega, \mathcal{F}, \Pr)$; G_n, D_n, and α_0 are then considered as functions $G_n(\cdot)$, $D_n(\cdot)$, and $\alpha_0(\cdot)$ from the sample space Ω to real values $\mathbb{R}$. From the discussions in Section 1.3.1, the effective intergeneration times $(G_\ell^\dagger(\omega))_{\ell=1,2,\ldots}$, the effective system delays $(D_\ell^\dagger(\omega))_{\ell=0,1,\ldots}$, the peak AoI values $(A_{\mathrm{peak},\ell}(\omega))_{\ell=1,2,\ldots}$, and the AoI values $(A_t(\omega))_{t\geq 0}$ for each sample-path $\omega \in \Omega$ are given in terms of $(G_n(\omega))_{n=1,2,\ldots}$, $(D_n(\omega))_{n=0,1,\ldots}$, and $\alpha_0(\omega)$.

Also, we see that Lemma 1.2 holds for each $\omega \in \Omega$. For stationary and ergodic systems, we can rewrite Lemma 1.2 as a relation of stationary probability distributions. More specifically, we make the following assumptions:

ASSUMPTION 1.3

(i) The joint process $(G_\ell^\dagger, D_\ell^\dagger)_{\ell=1,2,\ldots}$ of effective intergeneration times and effective delays is stationary and ergodic.
(ii) The effective generation rate is constant $\lambda^\dagger(\omega) = \lambda^\dagger$ almost surely (a.s.) with some $\lambda^\dagger \in (0, \infty)$.
(iii) The rate stability (1.21) holds a.s.
(iv) The AoI process $(A_t)_{t\geq 0}$ is stationary.

Remark 1.1 Assumption 1.3 (i) and (ii) have a little redundancy because the ergodicity of $(G_\ell^\dagger)_{\ell=1,2,\ldots}$ implies that $\lambda^\dagger(\omega)$ is constant a.s.

Recall that we have $A_{\mathrm{peak},\ell} = G_\ell^\dagger + D_\ell^\dagger$ as given in (1.3). The stationarity assumed in Assumption 1.3 thus implies the existence of generic random variables A, A_{peak}, and $D^\dagger$ with the same distributions as A_t, $A_{\mathrm{peak},\ell}$, and $D_\ell^\dagger$, respectively:

$$\begin{aligned}
\Pr(A_t \leq x) &= F_A(x), && x \geq 0,\ t \geq 0,\\
\Pr(A_{\mathrm{peak},\ell} \leq x) &= F_{A_{\mathrm{peak}}}(x), && x \geq 0,\ \ell = 1, 2, \ldots,\\
\Pr(D_\ell^\dagger \leq x) &= F_{D^\dagger}(x), && x \geq 0,\ \ell = 1, 2, \ldots.
\end{aligned}$$

For each sample-path $\omega \in \Omega$, let $F_A^\sharp(\omega, x)$, $F_{A_{\mathrm{peak}}}^\sharp(\omega, x)$, and $F_{D^\dagger}^\sharp(\omega, x)$ denote the AFDs of the AoI $(A_t(\omega))_{t\geq 0}$, the peak AoI $(A_{\mathrm{peak},\ell}(\omega))_{\ell=1,2,\ldots}$, and the effective system delay $(D_\ell^\dagger(\omega))_{\ell=0,1,\ldots}$, respectively (cf. (1.17), (1.18), and (1.19)).

The ergodicity of $(G_\ell^\dagger, D_\ell^\dagger)_{\ell=1,2,\ldots}$ (Assumption 1.3 (i)) implies that for $x \geq 0$,

$$F_{A_{\mathrm{peak}}}^\sharp(\omega, x) = F_{A_{\mathrm{peak}}}(x), \quad F_{D^\dagger}^\sharp(\omega, x) = F_{D^\dagger}(x), \quad \text{a.s.}$$

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

Therefore, we have, from Lemma 1.2,

$$F_A^{\sharp}(\omega, x) = \lambda^{\dagger} \int_0^x \big(F_{D^{\dagger}}(y) - F_{A_{\text{peak}}}(y)\big)\mathrm{d}y, \quad \text{a.s,}$$

which obviously implies

$$\mathrm{E}[F_A^{\sharp}(x)] = \lambda^{\dagger} \int_0^x \big(F_{D^{\dagger}}(y) - F_{A_{\text{peak}}}(y)\big)\mathrm{d}y.$$

On the other hand, we have, from the dominated convergence theorem,

$$\mathrm{E}[F_A^{\sharp}(x)] = \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathrm{E}[\mathbb{1}\{A_t \le x\}]\mathrm{d}t = F_A(x), \quad x \ge 0.$$

We thus conclude as in the following theorem:

THEOREM 1.4 *Under Assumption 1.3, the CDF of the stationary AoI distribution is given by*

$$F_A(x) = \lambda^{\dagger} \int_0^x \big(F_{D^{\dagger}}(y) - F_{A_{\text{peak}}}(y)\big)\mathrm{d}y, \quad x \ge 0.$$

Therefore, the AoI has an absolutely continuous distribution with density

$$f_A(x) = \lambda^{\dagger}\big(F_{D^{\dagger}}(x) - F_{A_{\text{peak}}}(x)\big), \quad x \ge 0. \tag{1.28}$$

Furthermore, the LST of the AoI is given by

$$f_A^*(s) = \lambda^{\dagger} \cdot \frac{f_{D^{\dagger}}^*(s) - f_{A_{\text{peak}}}^*(s)}{s}, \quad s > 0. \tag{1.29}$$

COROLLARY *The kth moment of the stationary AoI is given by*

$$\mathrm{E}[A^k] = \lambda^{\dagger} \cdot \frac{\mathrm{E}[(A_{\text{peak}})^{k+1}] - \mathrm{E}[(D^{\dagger})^{k+1}]}{k+1}, \tag{1.30}$$

provided that $\mathrm{E}[(A_{\text{peak}})^{k+1}] < \infty$.

Notice that the previously presented formula (1.6) for the mean AoI $\mathrm{E}[A]$ is reproduced by letting $k = 1$ in (1.30). In this sense, Theorem 1.4 can be regarded as a generalization of (1.6) to its distributional version. It is also worth noting that the expression (1.28) for the density function of the AoI distribution can be interpreted as *a level-crossing identity* (Brill & Posner 1977; Cohen 1977) for the AoI process. More specifically, for fixed $x \ge 0$, the density function $f_A(x)$ represents the mean number of upcrossings at level x of the AoI process per time unit because the AoI A_t linearly increases with slope one almost every t (cf. Figure 1.2). On the other hand, the right-hand side of (1.28) represents the mean number of downcrossings at level x per time unit, which is the product of the number of downcrossings (i.e., monitor updates) occurring per time unit and the probability of $\{A_{\text{peak},\ell} > x \ge D_{\ell}^{\dagger}\}$, which can be rewritten as

$$\Pr(A_{\text{peak},\ell} > x \geq D_\ell^\dagger) = \Pr(A_{\text{peak},\ell} > x) - \Pr(D_\ell^\dagger > x) = F_{D^\dagger}(x) - F_{A_{\text{peak}}}(x).$$

The formula (1.28) thus indicates that the mean numbers of upcrossings and downcrossings at level x occurring per time unit are equal.

Theorem 1.4 shows that the stationary distribution of the AoI A is given in terms of those of the peak AoI A_{peak} and the effective system delay $D^\dagger$. In many cases, characterizing the probability distributions of the peak AoI A_{peak} and the effective system delay $D^\dagger$ is quite easier than directly analyzing the AoI process. In the following section, we demonstrate how to apply Theorem 1.4 to AoI analysis, dealing with basic single-server queueing models as examples.

1.4 The AoI Distribution in FCFS and LCFS Single-Server Queues

In this section, we present an analysis of the AoI distribution for FCFS and last-come first-served (LCFS) single-server queues, which are described as follows. The sensor observes the state of a time-varying information source with independent and identically distributed (i.i.d.) generation intervals $(G_n)_{n=1,2,\ldots}$. The service system is represented as a single-server queue, where each update receives service with i.i.d. length of time. Let $(H_n)_{n=1,2,\ldots}$ denote the i.i.d. sequence of service times. We define G and H as generic random variables for intergeneration times and service times. We also define λ as the mean generation rate:

$$\lambda = \frac{1}{\mathrm{E}[G]}.$$

Let $\rho := \lambda\mathrm{E}[H]$ denote the traffic intensity. With Kendall's notation, this queueing model is denoted by GI/GI/1. We define $\widetilde{G}$ as a generic random variable for residual (equilibrium) intergeneration times, that is, the time to the next generation from a randomly chosen time instant. Similarly, we define $\widetilde{H}$ as a generic random variable for residual service times. The density functions and LSTs of $\widetilde{G}$ and $\widetilde{H}$ are given by

$$f_{\widetilde{G}}(x) = \frac{1 - F_G(x)}{\mathrm{E}[G]}, \qquad f_{\widetilde{G}}^*(s) = \frac{1 - f_G^*(s)}{s\mathrm{E}[G]}, \tag{1.31}$$

$$f_{\widetilde{H}}(x) = \frac{1 - F_H(x)}{\mathrm{E}[H]}, \qquad f_{\widetilde{H}}^*(s) = \frac{1 - f_H^*(s)}{s\mathrm{E}[H]}. \tag{1.32}$$

We consider two different service policies of the server: the FCFS and the preemptive LCFS service policies. Under the FCFS service policy, status updates are served in order of their arrivals, so that no overtaking of updates can occur. Under the preemptive LCFS service policy, on the other hand, the newest update is given priority: each update starts to receive service immediately after its arrival at the server, whereas its service is preempted when a newer update has arrived before the service completion. Recall that an update is said to be effective if it is not overtaken by other updates. It is readily seen that all updates are effective under the FCFS service policy, while there are noneffective updates under the preemptive LCFS service policy.

1.4.1 The Stationary AoI Distribution in FCFS Single-Server Queues

We first consider the FCFS case. Because all updates are effective in this case,

$$G_n^\dagger = G_n, \quad D_n^\dagger = D_n, \quad \lambda^\dagger = \lambda, \tag{1.33}$$

so that the peak AoI is given by (cf. (1.3))

$$A_{\text{peak},n} = D_n + G_n. \tag{1.34}$$

In order for Assumption 1.3 to be satisfied, we assume $\rho < 1$, which ensures the stability of the queueing system.

We note that D_n and G_n on the right-hand side of (1.34) are *dependent* in general: an update arriving after a long interval G_n tends to find less congested system than the time-average. For the FCFS queue, however, (1.34) can be rewritten in terms of independent random variables, using the well-known Lindley recursion for system delays $(D_n)_{n=0,1,\ldots}$: because the waiting time of the nth update equals $D_{n-1} - G_n$ if $D_{n-1} > G_n$ and otherwise it equals zero, we have

$$D_n = \max(0, D_{n-1} - G_n) + H_n, \quad n = 1, 2, \ldots.$$

(1.34) is then rewritten as

$$A_{\text{peak},n} = \max(G_n, D_{n-1}) + H_n, \quad n = 1, 2, \ldots. \tag{1.35}$$

Observe that G_n, D_{n-1}, and H_n are independent. Because we have

$$\begin{aligned} \Pr(\max(G_n, D_{n-1}) \le x) &= \Pr(G_n \le x, D_{n-1} \le x) \\ &= \Pr(G_n \le x) \Pr(D_{n-1} \le x), \end{aligned}$$

the relation (1.35) implies

$$F_{A_{\text{peak}}}(x) = \int_0^x F_G(x-y) F_D(x-y) \mathrm{d}F_H(y), \quad x \ge 0,$$

where D denotes a generic random variable for the stationary D_n. Therefore, we obtain the following result from Theorem 1.4:

THEOREM 1.5 *In the stationary FCFS GI/GI/1 queue, the probability density function of the AoI distribution is given by*

$$f_A(x) = \lambda \left(F_D(x) - \int_0^x F_G(x-y) F_D(x-y) \mathrm{d}F_H(y) \right). \tag{1.36}$$

Noting that λ, $F_G(\cdot)$, and $F_H(\cdot)$ are model parameters, we see from Theorem 1.5 that the AoI distribution is given in terms of the stationary delay distribution $F_D(\cdot)$. We can find expressions for the delay distribution $F_D(\cdot)$ for standard FCFS queueing systems in textbooks on the queueing theory (Asmussen 2003; Kleinrock 1975).

We first introduce a general result, assuming that service times follow a *phase-type* distribution with representation $(\boldsymbol{\gamma}, \boldsymbol{S})$:

$$F_H(x) = 1 - \boldsymbol{\gamma} \exp[\boldsymbol{S}x] \boldsymbol{e}, \quad f_H(x) = \boldsymbol{\gamma} \exp[\boldsymbol{S}x](-\boldsymbol{S}) \boldsymbol{e}, \tag{1.37}$$

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

where $\boldsymbol{e}$ denotes a column vector (with the same size as $\boldsymbol{S}$) whose elements are all equal to one. The queueing model considered is then denoted by GI/PH/1. We do not lose much generality with this assumption on service times because the set of phase-type distributions covers a fairly wide class of nonnegative probability distributions; in fact, it is known that the phase-type distributions form a dense subset in the set of all nonnegative probability distributions (Asmussen 2003, p. 84). Readers who are not familiar with phase-type distributions are advised to take a look at Appendix A.1, where we provide a brief introduction.

It is readily verified that for the phase-type service time distribution,

$$\mathrm{E}[H] = \boldsymbol{\gamma}(-\boldsymbol{S})^{-1}\boldsymbol{e}, \tag{1.38}$$

and the residual service time distribution defined in (1.32) is also of phase-type with representation $(\widetilde{\boldsymbol{\gamma}}, \boldsymbol{S})$, where

$$\widetilde{\boldsymbol{\gamma}} := \frac{\boldsymbol{\gamma}(-\boldsymbol{S})^{-1}}{\boldsymbol{\gamma}(-\boldsymbol{S})^{-1}\boldsymbol{e}}. \tag{1.39}$$

LEMMA 1.6 (Asmussen 1992) *Consider an FCFS GI/PH/1 queue, which has a general inter-arrival time distribution with CDF $F_G(x)$ and the phase-type service time distribution with representation $(\boldsymbol{\gamma}, \boldsymbol{S})$. The stationary system delay in this model follows a phase-type distribution with representation $(\boldsymbol{\gamma}, \boldsymbol{Q})$:*

$$F_D(x) = 1 - \boldsymbol{\gamma}\exp[\boldsymbol{Q}x]\boldsymbol{e}, \quad x \geq 0,$$

where $\boldsymbol{Q}$ is defined as

$$\boldsymbol{Q} = \boldsymbol{S} + (-\boldsymbol{S})\boldsymbol{e}\boldsymbol{\pi}_*, \tag{1.40}$$

with $\boldsymbol{\pi}_*$ defined as the limit $\boldsymbol{\pi}_* := \lim_{n\to\infty}\boldsymbol{\pi}_n$ of a sequence $(\boldsymbol{\pi}_n)_{n=0,1,\ldots}$ given by $\boldsymbol{\pi}_0 = \boldsymbol{0}$ and the following recursion:

$$\boldsymbol{\pi}_n = \boldsymbol{\gamma}\int_0^\infty \exp[(\boldsymbol{S} + (-\boldsymbol{S})\boldsymbol{e}\boldsymbol{\pi}_{n-1})y]\mathrm{d}F_G(y), \quad n = 1, 2, \ldots. \tag{1.41}$$

Remark 1.2 (Asmussen 2003, p. 241) $(\boldsymbol{\pi}_n)_{n=0,1,\ldots}$ is an (elementwise) nondecreasing sequence of subprobability vectors (i.e., $\boldsymbol{\pi}_n\boldsymbol{e} < 1$), and its limit $\boldsymbol{\pi}_*$ is also a subprobability vector, provided that the stability condition $\rho < 1$ holds.

We hereafter focus on the Poisson and constant generation policies discussed in Section 1.1.2, that is, we consider the FCFS M/PH/1 and D/PH/1 queues:

$$F_G(x) = 1 - e^{-\lambda x}, \quad x \geq 0, \quad \text{(M/PH/1)} \tag{1.42}$$

$$F_G(x) = \mathbb{1}\{x \geq \tau\}, \quad x \geq 0, \quad \text{(D/PH/1)}, \tag{1.43}$$

where $\tau := 1/\lambda$. In the M/PH/1 queue, $\boldsymbol{\pi}_*$ is given explicitly by

$$\boldsymbol{\pi}_* = \rho\widetilde{\boldsymbol{\gamma}}, \qquad \text{(M/PH/1)} \tag{1.44}$$

which can be verified with the following observation: substituting (1.42) into (1.41) and taking the limit $n \to \infty$, we have

$$\boldsymbol{\pi}_*(-\boldsymbol{S} + \lambda\boldsymbol{I} + \boldsymbol{S}\boldsymbol{e}\boldsymbol{\pi}_*) = \lambda\boldsymbol{\gamma}, \tag{1.45}$$

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

which is equivalent to

$$\boldsymbol{\pi}_*(-\boldsymbol{S}) = \lambda\boldsymbol{\gamma}, \tag{1.46}$$

because $\boldsymbol{\pi}_*(-\boldsymbol{S})\boldsymbol{e} = \lambda$ is obtained by post-multiplying both sides of (1.45) by $\boldsymbol{e}$ and rearranging terms. We thus obtain (1.44) from (1.38), (1.39), and (1.46). In the D/PH/1 queue, on the other hand, (1.41) is simplified as

$$\boldsymbol{\pi}_n = \boldsymbol{\gamma}\exp[(\boldsymbol{S} + (-\boldsymbol{S})\boldsymbol{e}\boldsymbol{\pi}_{n-1})\tau], \qquad \text{(D/PH/1)}$$

so that $\boldsymbol{\pi}_*$ is easily computed by iterations.

Furthermore, the formula (1.36) for the AoI distribution is simplified in the M/PH/1 and D/PH/1 queues:

THEOREM 1.7

(i) *In the stationary FCFS M/PH/1 queue, the density function and the CDF of the AoI are given by*

$$f_A(x) = \rho\boldsymbol{\gamma}_0\exp[\boldsymbol{B}_0x](-\boldsymbol{B}_0)\boldsymbol{e}_0 + \boldsymbol{\gamma}_1\exp[\boldsymbol{B}_1x](-\boldsymbol{B}_1)\boldsymbol{e}_1, \tag{1.47}$$

$$F_A(x) = 1 - \rho\boldsymbol{\gamma}_0\exp[\boldsymbol{B}_0x]\boldsymbol{e}_0 - \boldsymbol{\gamma}_1\exp[\boldsymbol{B}_1x]\boldsymbol{e}_1, \tag{1.48}$$

where $\boldsymbol{B}_0$ and $\boldsymbol{B}_1$ are defined as

$$\boldsymbol{B}_0 = \exp\left[\begin{pmatrix}\boldsymbol{Q} & (-\boldsymbol{Q})\boldsymbol{e}\tilde{\boldsymbol{\gamma}} \\ \boldsymbol{0} & \boldsymbol{S}\end{pmatrix}x\right],$$

$$\boldsymbol{B}_1 = \begin{pmatrix}\boldsymbol{Q} - \lambda\boldsymbol{I} & -(\boldsymbol{Q} - \lambda\boldsymbol{I})\boldsymbol{e}\boldsymbol{\gamma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S} & (-\boldsymbol{S})\boldsymbol{e} \\ \boldsymbol{0} & \boldsymbol{0} & -\lambda\end{pmatrix},$$

and $\boldsymbol{\gamma_0}$ (resp. $\boldsymbol{\gamma_1}$) denotes a row vector with the same size as $\boldsymbol{B}_0$ (resp. $\boldsymbol{B}_1$), which is expressed as

$$\boldsymbol{\gamma}_0 = [\boldsymbol{\gamma}\ \ \boldsymbol{0}], \qquad \boldsymbol{\gamma}_1 = [\boldsymbol{\gamma} - \boldsymbol{\pi}_*\ \ \boldsymbol{0}].$$

Also, $\boldsymbol{e}_0$ (resp. $\boldsymbol{e}_1$) denotes a column vector with the same size as $\boldsymbol{B}_0$ (resp. $\boldsymbol{B}_1$) whose elements are all equal to one.

(ii) *In the stationary FCFS D/PH/1 queue, the density function and the CDF of the AoI are given by*

$$f_A(x) = \begin{cases} \dfrac{1 - \boldsymbol{\gamma}\exp[\boldsymbol{Q}x]\boldsymbol{e}}{\tau}, & 0 \le x < \tau, \\[2ex] \dfrac{\boldsymbol{\gamma}(\boldsymbol{I} - \exp[\boldsymbol{Q}\tau])\exp[\boldsymbol{Q}(x-\tau)]\boldsymbol{e}}{\tau}, & x \ge \tau. \end{cases} \tag{1.49}$$

$$F_A(x) = \begin{cases} \dfrac{x - \boldsymbol{\gamma}(-\boldsymbol{Q})^{-1}\boldsymbol{e} + \boldsymbol{\gamma}\exp[\boldsymbol{Q}x](-\boldsymbol{Q})^{-1}\boldsymbol{e}}{\tau}, & 0 \le x < \tau, \\[2ex] 1 - \dfrac{\boldsymbol{\gamma}(\boldsymbol{I} - \exp[\boldsymbol{Q}\tau])\exp[\boldsymbol{Q}(x-\tau)](-\boldsymbol{Q})^{-1}\boldsymbol{e}}{\tau}, & x \ge \tau. \end{cases} \tag{1.50}$$

The derivation of Theorem 1.7 is detailed in Appendix A.2.

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

Using Theorem 1.7, the probability distribution of the AoI in M/PH/1 and D/PH/1 queues is easily computed. We provide a few numerical examples using the following special class of phase-type distributions, which is uniquely identified by its mean $\mathrm{E}[H]$ and coefficient of variation $\mathrm{Cv}[H]$ (the standard deviation divided by the mean):

Mixed Erlang Distribution $(0 < \mathrm{Cv}[H] < 1)$

$$f_H(x) = p\mu \cdot \frac{e^{-\mu x}(\mu x)^{k-1}}{(k-1)!} + (1-p)\mu \cdot \frac{e^{-\mu x}(\mu x)^k}{k!},$$

where

$$k = \lfloor 1/(\mathrm{Cv}[H])^2 \rfloor,$$
$$p = \frac{k+1}{1+(\mathrm{Cv}[H])^2}\left((\mathrm{Cv}[H])^2 - \sqrt{\frac{1-k(\mathrm{Cv}[H])^2}{k+1}}\right),$$
$$\mu = \frac{pk + (1-p)(k+1)}{\mathrm{E}[H]}.$$

This distribution is also represented as a phase-type distribution by letting $\boldsymbol{\gamma}$ and $\boldsymbol{T}$ be a row vector and a matrix of size $(k+1)$ given by

$$\boldsymbol{\alpha} = (1\ 0\ \ldots\ 0), \qquad \boldsymbol{T} = \begin{pmatrix} -\mu & \mu & 0 & \cdots & 0 & 0 & 0 \\ 0 & -\mu & \mu & \cdots & 0 & 0 & 0 \\ 0 & 0 & -\mu & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\mu & \mu & 0 \\ 0 & 0 & 0 & \cdots & 0 & -\mu & (1-p)\mu \\ 0 & 0 & 0 & \cdots & 0 & 0 & -\mu \end{pmatrix}.$$

Exponential Distribution $(\mathrm{Cv}[H] = 1)$

$$f_H(x) = \mu e^{-\mu x}.$$

This is a phase-type distribution with $\boldsymbol{\gamma} = 1$ and $\boldsymbol{S} = -\mu$, and we have $\mu = 1/\mathrm{E}[H]$. This case corresponds to the M/M/1 and D/M/1 queues.

Hyper-Exponential Distribution with Balanced Means $(\mathrm{Cv}[H] > 1)$

$$f_H(x) = p\mu_1 e^{-\mu_1 x} + (1-p)\mu_2 e^{-\mu_2 x},$$

where

$$p = \frac{1}{2}\left(1 + \sqrt{\frac{(\mathrm{Cv}[H])^2 - 1}{(\mathrm{Cv}[H])^2 + 1}}\right), \quad \mu_1 = \frac{2p}{\mathrm{E}[H]}, \quad \mu_2 = \frac{2(1-p)}{\mathrm{E}[H]}.$$

This is a phase-type distribution with

$$\boldsymbol{\gamma} = (p \quad 1-p), \qquad \boldsymbol{S} = \begin{pmatrix} -\mu_1 & 0 \\ 0 & -\mu_2 \end{pmatrix}.$$

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

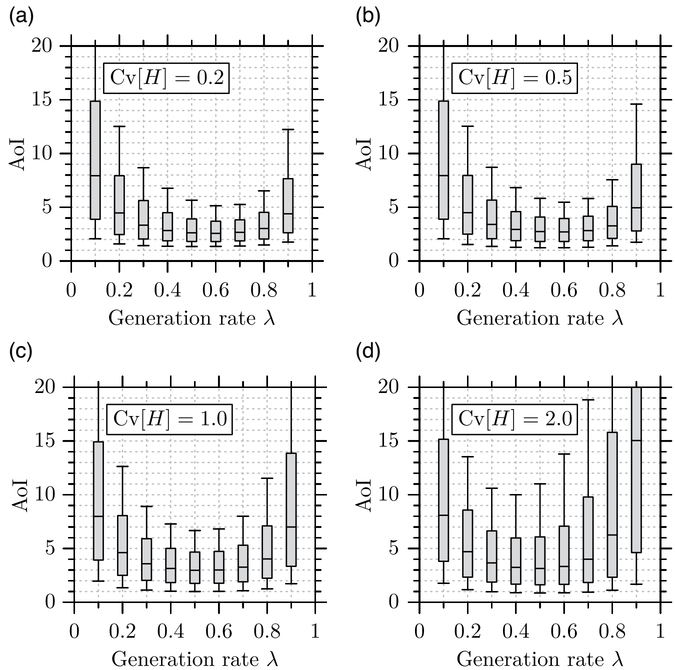


Figure 1.7 Boxplots of the AoI distribution in the FCFS M/PH/1 queue. The whiskers represent 10 and 90 percentiles.

We set $\mathrm{E}[H] = 1$ throughout. Figure 1.7 shows boxplots of the AoI distribution in the M/PH/1 queue for several values of the coefficient of variation $\mathrm{Cv}[H]$ of service times and the generation rate λ. As previously mentioned, the case of $\mathrm{Cv}[H] = 1.0$ refers to the M/M/1 queue, whose mean AoI was examined with Figure 1.4. We observe that using too small or too large generation rate λ leads to a significant increase in the variability of the AoI as well as the increase in the median value. We also observe that larger variability of service times leads to a significant increase in the AoI percentiles: its effect on the AoI distribution is prominent, particularly for large values of λ, because the system delay has a dominant impact on the AoI in that region, as discussed in Section 1.1. Figure 1.8 shows similar boxplots of the AoI distribution in the D/PH/1 queue. We see that the same observations as those for the M/M/1 queue apply regarding the impacts of the generation rate and the service-time variability on the AoI distribution. On the other hand, from Figures 1.7 and 1.8, we see that the variability of the AoI is reduced drastically by using constant generation intervals instead of exponential generation intervals. This again highlights the superiority of using constant generation intervals, which we observed at the end of Section 1.1.2.

1.4.2 The Stationary AoI Distribution in Preemptive LCFS Single-Server Queues

We next consider the case of preemptive LCFS service policy. In this case, an update immediately starts to receive service just after its generation. The update finishes the

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

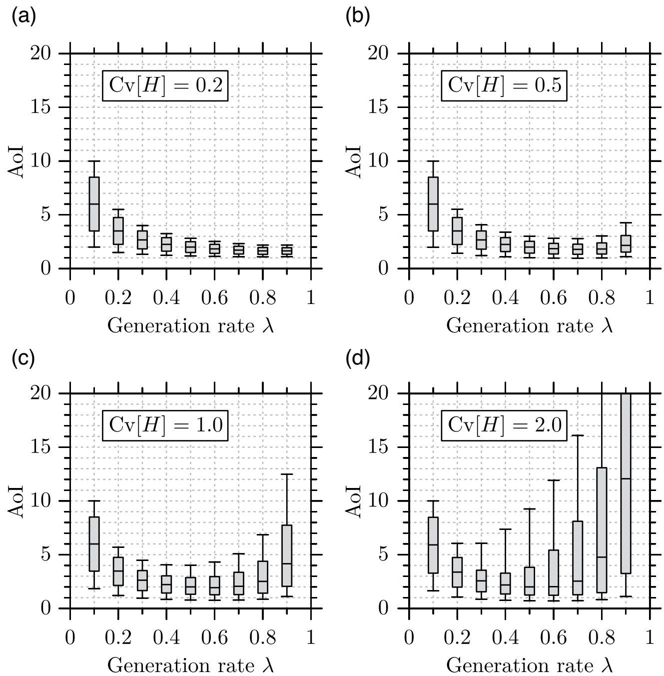


Figure 1.8 Boxplots of the AoI distribution in the FCFS D/PH/1 queue. The whiskers represent 10 and 90 percentiles.

service without being overtaken (i.e., becomes an effective update) if the next update is not generated until the end of the service. If the generation of the next update occurs before the service completion, on the other hand, the update is overtaken and becomes noneffective. Therefore, the probability that an update becomes noneffective is given by

$$\zeta := \Pr(G < H).$$

If $\zeta = 1$, then updates cannot finish receiving service with probability one, so that the AoI continues to increase all the time and never decreases a.s. If $\zeta = 0$, on the other hand, updates become effective with probability one, so that the model reduces to the FCFS case. We thus assume $0 < \zeta < 1$ hereafter. Also, we assume $\Pr(G = H) = 0$ for simplicity. Similarly to the FCFS case, we first derive a general result for the GI/GI/1 queue and then specialize it to the M/PH/1 and D/PH/1 queues.

We use the notation $[Y \mid \mathcal{E}]$ to represent a random variable Y conditioned on an event $\mathcal{E}$. From the previous discussion, we see that the effective system delay $D^\dagger$ has the same distribution as the conditional random variable $H_{<G} := [H \mid H < G]$:

$$D^\dagger \overset{\mathrm{d}}{=} H_{<G}, \tag{1.51}$$

where $\overset{\mathrm{d}}{=}$ stands for the equality in distribution. Recall that the effective intergeneration time $G_\ell^\dagger$ is given by the sum of intergeneration times of the form (1.15). In the current setting, we have that the terms on the right-hand side of (1.15) are characterized as

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

$$G_{n+1} \stackrel{\mathrm{d}}{=} G_{>H}, \quad G_{n+i} \stackrel{\mathrm{d}}{=} G_{<H}, \quad i = 2, 3, \ldots, k+1,$$

where we define $G_{>H} = [G \mid G > H]$ and $G_{<H} = [G \mid G < H]$. We can also verify that these intergeneration times $G_{n+1}, G_{n+2}, \ldots, G_{n+k+1}$ are independent. Furthermore, the number k of intermediate noneffective updates follows a geometric distribution with success probability $\Pr(G > H) = 1 - \zeta$; let K denote this discrete random variable:

$$\Pr(K = k) = \zeta^k (1 - \zeta), \quad k = 0, 1, \ldots. \tag{1.52}$$

The distribution of the effective intergeneration time is then characterized as

$$G_\ell^\dagger \stackrel{\mathrm{d}}{=} G_{>H} + \sum_{i=1}^{K} G_{<H}^{[i]},$$

where $(G_{<H}^{[i]})_{i=1,2,\ldots}$ denotes an i.i.d. sequence of random variables with the same distribution as $G_{<H}$. From (1.3) and (1.51), we then have that

$$A_{\mathrm{peak},\ell} \stackrel{\mathrm{d}}{=} G_{>H} + \sum_{i=1}^{K} G_{<H}^{[i]} + H_{<G}, \tag{1.53}$$

and that all random variables on the right-hand side of this equation are independent; note that $D_\ell^\dagger$ is dependent on *the next* intergeneration time G_{n+k+1} (as $D_\ell^\dagger < G_{n+k+1}$ should hold for this update to be effective), whereas it is independent of the preceding intergeneration times $G_{n+1}, G_{n+2} \ldots, G_{n+k}$.

As the expression (1.53) for the peak AoI contains the geometric random sum of i.i.d. random variables, it is easier to consider its LST instead of directly dealing with its CDF. More specifically, the LST of the geometric random sum is expressed as

$$\mathrm{E}\left[\exp\left[-s \sum_{i=1}^{K} G_{<H}^{[i]}\right]\right] = \sum_{k=0}^{\infty} \zeta^k (1 - \zeta) \{f_{G_{<H}}^*(s)\}^k = \frac{1 - \zeta}{1 - \zeta f_{G_{<H}}^*(s)}.$$

Therefore, the LST of the peak AoI is given by

$$f_{A_{\mathrm{peak}}}^*(s) = f_{G_{>H}}^*(s) \cdot \frac{1 - \zeta}{1 - \zeta f_{G_{<H}}^*(s)} \cdot f_{H_{<G}}^*(s).$$

Substituting this equation, $\lambda^\dagger = \lambda(1 - \zeta)$, and $f_{D^\dagger}^*(s) = f_{H_{<G}}^*(s)$ (cf. (1.51)) into (1.29), and rearranging terms using

$$f_G^*(s) = \zeta f_{G_{<H}}^*(s) + (1 - \zeta) f_{G_{>H}}^*(s),$$

we obtain the following result:

THEOREM 1.8 *In the preemptive LCFS GI/GI/1 queue, the LST of the stationary AoI distribution is given by*

$$f_A^*(s) = f_{\widetilde{G}}^*(s) \cdot \frac{(1 - \zeta) f_{H_{<G}}^*(s)}{1 - \zeta f_{G_{<H}}^*(s)}, \tag{1.54}$$

where $\widetilde{G}$ is defined as in (1.31).

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

Furthermore, we can interpret (1.54) as follows:

COROLLARY *In the preemptive LCFS GI/GI/1 queue, the stationary AoI has the same distribution as the sum of two independent random variables,*

$$A = \widetilde{G} + Z,$$

where Z is defined as

$$Z = \sum_{n=1}^{\widehat{K}} \min(G_n, H_n), \tag{1.55}$$

$$\widehat{K} = \min\{n = 1, 2, \ldots; H_n < G_n\}, \tag{1.56}$$

in terms of the i.i.d. sequences $(G_n)_{n=1,2,\ldots}$ *and* $(H_n)_{n=1,2,\ldots}$ *of intergeneration and service times.*

Remark 1.3 By definition, we have $\widehat{K} \stackrel{\mathrm{d}}{=} K + 1$ (cf. (1.52)).

We then focus on the M/PH/1 and D/PH/1 queues, where the service time distribution is given by (1.37). We can readily verify from (1.31) that the residual intergeneration time $\widetilde{G}$ has simple characterizations in these cases:

$$f_{\widetilde{G}}(x) = \lambda e^{-\lambda x}, \quad x \geq 0, \qquad \text{(M/PH/1)} \tag{1.57}$$

$$f_{\widetilde{G}}(x) = \frac{1}{\tau} \cdot \mathbb{1}\{x \leq \tau\}, \quad x \geq 0, \qquad \text{(D/PH/1)}, \tag{1.58}$$

that is, $\widetilde{G}$ follows an exponential (resp. a uniform) distribution in the M/PH/1 (resp. the D/PH/1) queue. Therefore, it is sufficient to consider the distribution of Z defined in (1.55).

In the M/PH/1 queue, Z follows a phase-type distribution with representation $(\boldsymbol{\gamma}, \boldsymbol{S} - \lambda \boldsymbol{I} + \lambda \boldsymbol{e}\boldsymbol{\gamma})$, that is,

$$F_Z(x) = 1 - \boldsymbol{\gamma} \exp[(\boldsymbol{S} - \lambda \boldsymbol{I} + \lambda \boldsymbol{e}\boldsymbol{\gamma})x]\boldsymbol{e}. \quad \text{(M/PH/1)} \tag{1.59}$$

This fact can be verified with the following observation. Note that $\min(G_n, H_n)$ in the M/PH/1 queue is equal to the absorption time of a Markov chain with transition rate matrix $\boldsymbol{S} - \lambda \boldsymbol{I}$ and initial-state distribution $\boldsymbol{\gamma}$. This Markov chain transitions to its absorbing state according to a rate vector $(-\boldsymbol{S})\boldsymbol{e} + \lambda \boldsymbol{e}$, where the first (resp. second) term corresponds to the event $\min(G_n, H_n) = H_n$ (resp. $\min(G_n, H_n) = G_n$). Therefore, we can obtain a realization of Z with the following procedure: (i) start this Markov chain with the initial-state distribution $\boldsymbol{\gamma}$, (ii) whenever an absorption with the rate vector $\lambda \boldsymbol{e}$ occurs, restart the Markov chain with the initial-state distribution $\boldsymbol{\gamma}$, and (iii) terminate the process when an absorption with the rate vector $(-\boldsymbol{S})\boldsymbol{e}$ occurs. The whole process of this procedure is then represented as a Markov chain with initial-state vector $\boldsymbol{\gamma}$ and transition rate matrix $\boldsymbol{S} - \lambda \boldsymbol{I} + \lambda \boldsymbol{e}\boldsymbol{\gamma}$, which proves the preceding claim.

In the D/PH/1 queue, on the other hand, we have $G_{<H} = \tau$ and $H_{<G} \in [0, \tau)$ with probability one. Also, we have

$$\zeta = \boldsymbol{\gamma} \exp[\boldsymbol{S}\tau]\boldsymbol{e}. \quad \text{(D/PH/1)} \tag{1.60}$$

It is then readily verified that for $k = 0, 1, \ldots,$

$$F_Z(x) = 1 - \zeta^k \boldsymbol{\gamma} \exp\big[\boldsymbol{S}(x - k\tau)\big]\boldsymbol{e}, \quad x \in [k\tau, (k+1)\tau). \quad \text{(D/PH/1)} \tag{1.61}$$

Therefore, we obtain the AoI distribution in the M/PH/1 and D/PH/1 queues from Corollary 1.4.2, (1.57), (1.58), (1.59), and (1.61):

THEOREM 1.9

(i) In the stationary preemptive LCFS M/PH/1 queue, the density function and the CDF of the AoI are given by

$$f_A(x) = \boldsymbol{\gamma} \exp[\boldsymbol{C}x](-\boldsymbol{C})\boldsymbol{e}, \quad x \geq 0,$$
$$F_A(x) = 1 - \boldsymbol{\gamma} \exp[\boldsymbol{C}x]\boldsymbol{e}, \quad x \geq 0,$$

where $\boldsymbol{C}$ *is defined as*

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{S} - \lambda \boldsymbol{I} + \lambda \boldsymbol{e}\boldsymbol{\gamma} & (-\boldsymbol{S})\boldsymbol{e} \\ \boldsymbol{0} & -\lambda \end{pmatrix}.$$

(ii) In the stationary preemptive LCFS D/PH/1 queue, the density function and the CDF of the AoI are given as follows: for $x \in [0, \tau)$*,*

$$f_A(x) = \frac{1 - \boldsymbol{\gamma} \exp[\boldsymbol{S}x]\boldsymbol{e}}{\tau},$$
$$F_A(x) = \frac{x - \boldsymbol{\gamma}(\boldsymbol{I} - \exp[\boldsymbol{S}x])(-\boldsymbol{S})^{-1}\boldsymbol{e}}{\tau},$$

and for $x \in [k\tau, (k+1)\tau)$ $(k = 1, 2, \ldots)$*,*

$$f_A(x) = \frac{\zeta^{k-1}(1-\zeta)}{\tau} \cdot \boldsymbol{\gamma} \exp[\boldsymbol{S}(x - k\tau)]\boldsymbol{e},$$
$$F_A(x) = 1 - \frac{\zeta^{k-1}\boldsymbol{\gamma}}{\tau}\Big(\zeta \boldsymbol{I} - \exp[\boldsymbol{S}\tau] + (1-\zeta)\exp\big[\boldsymbol{S}(x - k\tau)\big]\Big)(-\boldsymbol{S})^{-1}\boldsymbol{e},$$

where ζ *is given by (1.60).*

Finally, we present numerical examples employing the same class of service time distributions as those in the previous subsection, that is, the mixed Erlang and balanced hyper-exponential distributions with mean $\mathrm{E}[H] = 1$, which is uniquely identified by the coefficient of variation $\mathrm{Cv}[H]$. Figures 1.9 and 1.10 show comparisons of the FCFS and preemptive LCFS service policies in terms of the AoI distribution for M/PH/1 and D/PH/1 queues. As we previously observed in Figures 1.7 and 1.8, the large variability of service times has a significant negative impact on the AoI performance under the FCFS service policy. From Figures 1.9 and 1.10, however, we observe that the AoI performance is not negatively impacted under the preemptive LCFS service policy. Moreover, quite the contrary to the FCFS case, we find that *under the preemptive LCFS service policy, the AoI even tends to be smaller when service times have larger variability*. Intuitively, this phenomenon is due to the fact that the larger the variability

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

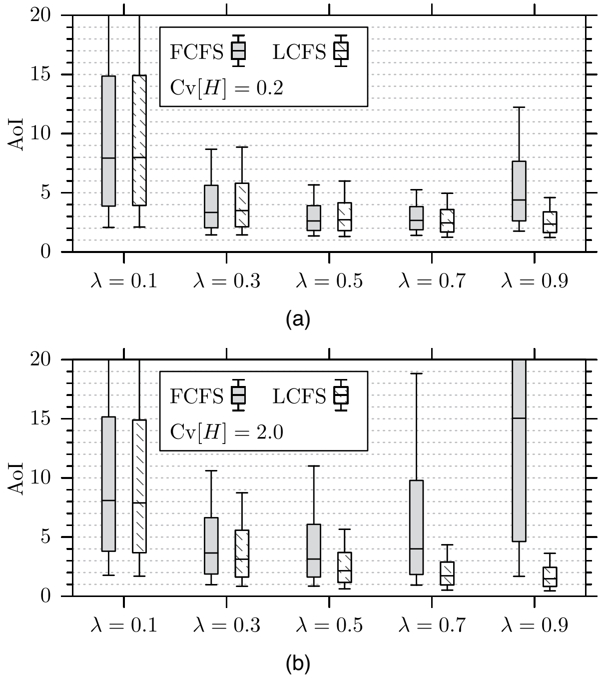


Figure 1.9 Comparison of the AoI distribution in FCFS and LCFS M/PH/1 queues.

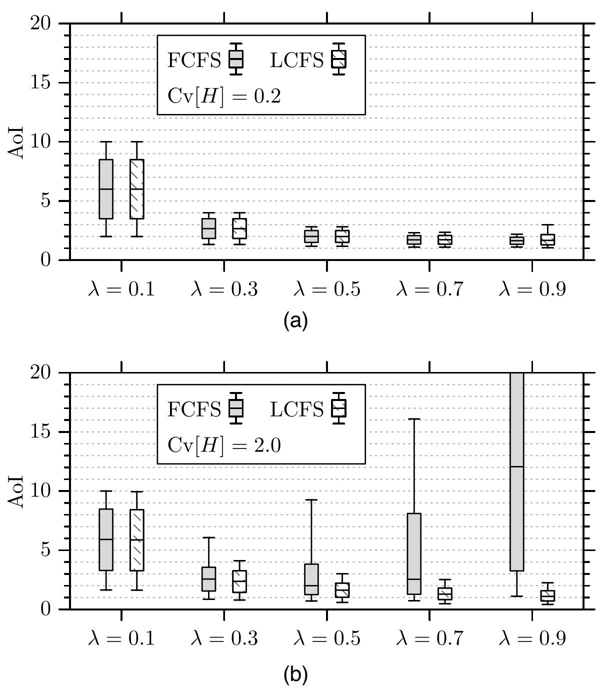


Figure 1.10 Comparison of the AoI distribution in FCFS and LCFS D/PH/1 queues.

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

of the service times, the more frequently service times take substantially smaller values than their mean. Updates with such small service times can update the monitor just after their generations under the preemptive LCFS service policy, which contributes to the timeliness of information updates.

While the preemptive LCFS service policy is quite efficient for highly variable service times, it should be noted that this policy *does not* always achieve superior performance compared to the FCFS service policy. For the case of $\mathrm{Cv}[H] = 0.2$ in Figure 1.9, we observe that the preemptive LCFS service policy has slightly worse AoI performance (in terms of the percentiles) than the FCFS service policy, except for large values of λ. This is because, when service times have small variability, the preemptive LCFS service policy may cause inefficient interruptions of service, resulting in a long update interval.

1.5 Bibliographical Notes

Section 1.1 The graphical analysis of the mean AoI and the queueing modeling of monitoring systems are due to the seminal work (Kaul, Yates, & Gruteser 2012b) on the AoI analysis, where the formulas for the mean AoI in the FCFS M/M/1 and D/M/1 queues are given. The expression (1.8) for the mean AoI focusing on the inter-departure time is given by Costa, Codreanu, and Ephremides (2016), where the authors also introduce the concept of the peak AoI A_{peak} as an alternative freshness metric.

Section 1.2 The nonlinear AoI cost function (1.12) is introduced by Sun, Uysal-Biyikoglu, Yates, Koksal, and Shroff (2017), where the optimal sampling policy minimizing the mean nonlinear AoI cost is discussed. Several expressions for the mean nonlinear AoI cost are also derived by Kosta, Pappas, Ephremides, and Angelakis (2020a) assuming specific cost functions. The effect of the AoI on the monitoring accuracy is first considered by Costa, Valentin, and Ephremides (2015) for a discrete-time two-state Markov chain, where the monitoring accuracy is defined as the probability that the actual and displayed states coincide. The monitoring accuracy for a continuous-time Markov chain is also analyzed by Inoue and Takine (2019). Sun, Polyanskiy, and Uysal (2020) show that if the information source is represented as a Wiener process and the AoI process is not affected by its state, then the mean AoI equals the mean squared error of the displayed information. Sun and Cyr (2018) also propose the use of the mutual information in quantifying the monitoring accuracy.

Sections 1.3 and 1.4 Most of the discussions in these sections are based on those given by Inoue, Masuyama, Takine, and Tanaka (2019), where the formula (1.4) for the AoI distribution is derived under a more general setting. In fact, this formula follows from the observation that the stochastic properties of the stationary AoI are characterized in terms of those of AoI values immediately before and after an update, so that the assumptions on underlying mechanisms (i.e., sampling intervals $(G_n)_{n=0,1,\ldots}$ and system delays $(D_n)_{n=0,1,\ldots}$) are not essential. Theorems 1.7 and 1.9 for the AoI distribution in queues with phase-type service time distributions are new, although most of those are specializations of more general results derived by Inoue et al. (2019). We have

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

not included discussions on LCFS service policies without service preemption for the brevity of explanation. The analysis of the AoI distribution for non-preemptive LCFS queues can be found in the papers (Inoue et al. 2019) and Champati, Al-Zubaidy, and Gross (2019), which are extensions of earlier studies on the mean AoI analysis (Kaul, Yates, and Gruteser 2012a; Costa et al. 2016).

Further Studies on the AoI Distribution The general formula (1.4) serves as a starting point of the analysis of the distributional properties of the AoI. For instance, its applications to a single-server queue with packet deadlines (Inoue 2018) and infinite-server queues (Inoue 2020) are reported. While this formula is derived assuming continuous-time systems, a discrete-time analog of this formula is also considered by Kosta, Pappas, Ephremides, and Angelakis (2020b). Furthermore, Jiang and Miyoshi (2020) give a generalization of (1.4) to *the joint distribution* of the AoI values considering multiple source-monitor pairs.

Appendix

A.1 A Brief Summary of Phase-Type Distributions

Consider an absorbing Markov chain $(X_t)_{t\geq 0}$ with finite states. Generally, states in an absorbing Markov chain are classified into transient states and absorbing states, and the Markov chain eventually reaches an absorbing state after repeating transitions among transient states. The phase-type distribution is a class of probability distributions on $[0,\infty)$, which can be formulated as the absorption time T (i.e., the first passage time to a set of absorbing states) in an absorbing Markov chain. In the context of the phase-type distribution, transient states are called phases. In this appendix, the absorbing Markov chain $(X_t)_{t\geq 0}$ is assumed to have M ($M \geq 1$) phases and let $\mathcal{M} = \{1,2,\ldots,M\}$.

In general, the absorbing Markov chain $(X_t)_{t\geq 0}$ is defined by the M-dimensional initial state vector $\boldsymbol{\alpha}$ and $M \times M$ matrix $\boldsymbol{Q}$, where the ith ($i \in \mathcal{M}$) element of $\boldsymbol{\alpha}$ is given by $\Pr(X_0 = i)$ and the (i,j)th ($i,j \in \mathcal{M}$, $i \neq j$) off-diagonal element $q_{i,j}$ of $\boldsymbol{Q}$ is given by the transition rate from state i to state j, that is,

$$q_{i,j} = \lim_{\Delta t\to 0} \frac{\Pr(X_{t+\Delta t} = j \mid X_t = i)}{\Delta t}, \quad i \neq j,$$

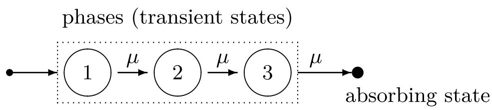
and the ith ($i \in \mathcal{M}$) diagonal element $q_{i,i}$ is given by

$$q_{i,i} = -\sum_{j\in M\setminus\{i\}} q_{i,j}.$$

We assume that $\boldsymbol{Q} + (-\boldsymbol{Q})\boldsymbol{e}\boldsymbol{\alpha}$ is irreducible. For a given $(\boldsymbol{\alpha},\boldsymbol{Q})$, the probability distribution function $F_T(x) = \Pr(T \leq x)$ of the absorption time T is given by

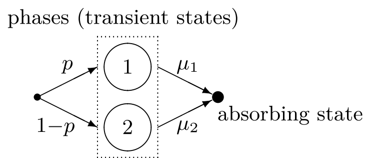
$$F_T(x) = 1 - \boldsymbol{\alpha}\exp[\boldsymbol{Q}x]\boldsymbol{e}, \quad x \geq 0,$$

where $\boldsymbol{e}$ denotes a column vector with an appropriate dimension, whose elements are all equal to one. Note here that $F_T(0) = \Pr(T = 0) = 1 - \boldsymbol{\alpha}\boldsymbol{e} \geq 0$. For simplicity,



$$\boldsymbol{\alpha} = (1\ 0\ 0), \quad \boldsymbol{Q} = \begin{pmatrix} -\mu & \mu & 0 \\ 0 & -\mu & \mu \\ 0 & 0 & -\mu \end{pmatrix}$$

Figure A.1 A three-stage Erlang distribution.



$$\boldsymbol{\alpha} = (p\ 1-p), \quad \boldsymbol{Q} = \begin{pmatrix} -\mu_1 & 0 \\ 0 & -\mu_2 \end{pmatrix}$$

Figure A.2 A two-stage hyper-exponential distribution.

we assume $F_T(0) = 0$, that is, $\boldsymbol{\alpha e} = 1$. The probability density function $f_T(x)$ is then given by

$$f_T(x) = \boldsymbol{\alpha} \exp[\boldsymbol{Q}x](-\boldsymbol{Q})\boldsymbol{e}, \quad x \geq 0,$$

and the nth ($n = 1, 2, \ldots$) moment $\mathrm{E}[T^n]$ is given by

$$\mathrm{E}[T^n] = \int_0^\infty x^n f(x)\mathrm{d}x = n!\boldsymbol{\alpha}\big[(-\boldsymbol{Q})^{-1}\big]^n \boldsymbol{e}.$$

Figures A.1 and A.2 show phase diagrams for a three-stage Erlang distribution and a two-stage hyper-exponential distribution.

We define $\widetilde{\boldsymbol{\alpha}}$ as a $1 \times M$ probability vector given by

$$\widetilde{\boldsymbol{\alpha}} = \frac{\boldsymbol{\alpha}(-\boldsymbol{Q})^{-1}}{\boldsymbol{\alpha}(-\boldsymbol{Q})^{-1}\boldsymbol{e}}.$$

Note that $\widetilde{\boldsymbol{\alpha}}$ satisfies $\widetilde{\boldsymbol{\alpha}}[\boldsymbol{Q} + (-\boldsymbol{Q})\boldsymbol{e\alpha}] = \boldsymbol{0}$. The probability distribution function $F_{\widetilde{T}}(x)$ of the equilibrium random variable $\widetilde{T}$ for T is then given by

$$F_{\widetilde{T}}(x) = 1 - \widetilde{\boldsymbol{\alpha}} \exp[\boldsymbol{Q}x]\boldsymbol{e}, \quad x \geq 0.$$

The phase-type distribution is close under the convolution, minimum, and maximum operations. Let T_i ($i = 1, 2$) denote an independent phase-type random variable with representation $(\boldsymbol{\alpha}_i, \boldsymbol{Q}_i)$. The probability distribution function of $T_1 + T_2$ is then given by

$$\Pr(T_1 + T_2 \leq x) = 1 - \boldsymbol{\alpha}_{\mathrm{sum}} \exp[\boldsymbol{Q}_{\mathrm{sum}}x]\boldsymbol{e}, \quad x \geq 0,$$

where

$$\boldsymbol{\alpha}_{\mathrm{sum}} = \begin{pmatrix} \boldsymbol{\alpha}_1 & \boldsymbol{0} \end{pmatrix}, \qquad \boldsymbol{Q}_{\mathrm{sum}} = \begin{pmatrix} \boldsymbol{Q}_1 & (-\boldsymbol{Q}_1)\boldsymbol{e\alpha}_2 \\ \boldsymbol{O} & \boldsymbol{Q}_2 \end{pmatrix}.$$

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

We can also obtain the phase-type distribution functions of $\min(T_1, T_2)$ and $\max(T_1, T_2)$ using Kronecker product/sum. In particular, if T_2 follows an exponential distribution with parameter λ (i.e., the phase-type distribution with representation $(1, -\lambda)$), we have

$$\Pr(\min(T_1, T_2) \le x) = 1 - \boldsymbol{\alpha}_1 \exp[(\boldsymbol{Q}_1 - \lambda \boldsymbol{I})x]\boldsymbol{e}, \quad x \ge 0,$$
$$\Pr(\max(T_1, T_2) \le x) = 1 - \boldsymbol{\alpha}_{\max} \exp[\boldsymbol{Q}_{\max} x]\boldsymbol{e}, \quad x \ge 0,$$

where

$$\boldsymbol{\alpha}_{\max} = \begin{pmatrix} \boldsymbol{\alpha}_1 & \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}, \qquad \boldsymbol{Q}_{\max} = \begin{pmatrix} \boldsymbol{Q}_1 - \lambda \boldsymbol{I} & \lambda \boldsymbol{I} & (-\boldsymbol{Q}_1)\boldsymbol{e} \\ \boldsymbol{O} & \boldsymbol{Q}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & -\lambda \end{pmatrix}.$$

Another advantage of the phase-type distribution is computational feasibility, that is, the uniformization technique is applicable. For a phase-type distribution with representation $(\boldsymbol{\alpha}, \boldsymbol{Q})$, we define θ as

$$\theta = \max_{i \in \mathcal{M}} |q_{i,i}|.$$

It can be readily verified that

$$\exp[\boldsymbol{Q}x] = e^{-\theta x} \exp[(\boldsymbol{I} + \theta^{-1}\boldsymbol{Q})\theta x] = \sum_{n=0}^{\infty} e^{-\theta x} \frac{(\theta x)^n}{n!} \boldsymbol{P}^n,$$

where $\boldsymbol{P} = \boldsymbol{I} + \theta^{-1}\boldsymbol{Q}$. Note here that $\boldsymbol{P}$ can be regarded as a defective transition probability matrix because $\boldsymbol{I} + \theta^{-1}\boldsymbol{Q} \ge \boldsymbol{O}$ and $(\boldsymbol{I} + \theta^{-1}\boldsymbol{Q})\boldsymbol{e} \le \boldsymbol{e}$. The technique for expressing $\exp[\boldsymbol{Q}t]$ in terms of $\boldsymbol{P}$ is called uniformization.

The probability distribution of the phase-type random variable T with representation $(\boldsymbol{\alpha}, \boldsymbol{Q})$ may be computed as follows. For $x \ge 0$, let $\overline{F}_T(x) = \Pr(T > x) = \boldsymbol{\alpha} \exp[\boldsymbol{Q}x]\boldsymbol{e}$. The following procedure yields an approximation $\overline{F}_T^{\text{approx}}(x)$ to $\overline{F}_T(x)$, which satisfies

$$0 < \overline{F}_T(x) - \overline{F}_T^{\text{approx}}(x) \le \epsilon,$$

for a predefined error bound ϵ $(0 < \epsilon \ll 1)$.

1. Find a positive integer N^* such that

$$\sum_{n=0}^{N^*} e^{-\theta x} \frac{(\theta x)^n}{n!} \ge 1 - \epsilon.$$

2. Let

$$\boldsymbol{y}_{N^*} = \frac{\theta x}{N^*} \boldsymbol{P}\boldsymbol{e},$$

and compute

$$\boldsymbol{y}_n = \frac{\theta x}{n} \boldsymbol{P}(\boldsymbol{e} + \boldsymbol{y}_{n+1}),$$

recursively in descending order from $n = N^* - 1$ to 1.

3. Let $\overline{F}_T^{\text{approx}}(x) = \exp(-\theta x)\boldsymbol{\alpha}(\boldsymbol{e} + \boldsymbol{y}_1)$.

https://doi.org/10.1017/9781108943321.001 Published online by Cambridge University Press

A.2 Derivation of Theorem 1.7

We first consider (i). Because the intergeneration time G and the system delay D are absolutely continuous in the M/PH/1 queue, we have that $\max(G, D)$ is also absolutely continuous with density:

$$\begin{aligned}
f_{\max(G,D)}(x) &= f_G(x)F_D(x) + F_G(x)f_D(x) \\
&= \lambda e^{-\lambda x}F_D(x) + (1 - e^{-\lambda x})f_D(x) \\
&= f_D(x) + e^{-\lambda x}(\lambda F_D(x) - f_D(x)) \\
&= f_D(x) - \frac{\mathrm{d}}{\mathrm{d}x}\left[F_D(x)e^{-\lambda x}\right].
\end{aligned}$$

We then rewrite (1.36) as

$$\begin{aligned}
f_A(x) &= \lambda\left\{\int_0^x f_D(y)\mathrm{d}y - \int_0^x \left(f_D(y) - \frac{\mathrm{d}}{\mathrm{d}y}\left[F_D(x)e^{-\lambda y}\right]\right) F_H(x-y)\mathrm{d}y\right\} \\
&= \rho \int_0^x f_D(y) \cdot \frac{1 - F_H(x-y)}{\mathrm{E}[H]}\mathrm{d}y + \lambda \int_0^x \frac{\mathrm{d}}{\mathrm{d}y}\left[F_D(y)e^{-\lambda y}\right] F_H(x-y)\mathrm{d}y \\
&= \rho f_D * f_{\widetilde{H}}(x) + \lambda \int_0^x F_D(y)e^{-\lambda y} f_H(x-y)\mathrm{d}y \\
&= \rho f_D * f_{\widetilde{H}}(x) + \int_{y=0}^x \left(\int_{t=0}^y f_D(t)e^{-\lambda t} \cdot \lambda e^{-\lambda(y-t)}\mathrm{d}t\right) f_H(x-y)\mathrm{d}y \\
&= \rho f_D * f_{\widetilde{H}}(x) + (1-\rho) f_{D_{<G}} * f_G * f_H(x),
\end{aligned} \tag{A.62}$$

where $D_{<G} := [D \mid D < G]$ denotes the system delay D conditioned to be smaller than an intergeneration time G. The density function of $D_{<G}$ is given by

$$\begin{aligned}
f_{D_{<G}}(x) &= \frac{f_D(x)e^{-\lambda x}}{\Pr(D < G)} \\
&= \frac{f_D(x)e^{-\lambda x}}{1-\rho} \\
&= \frac{\boldsymbol{\gamma} \exp[(\boldsymbol{Q} - \lambda \boldsymbol{I})x](-\boldsymbol{Q})\boldsymbol{e}}{1-\rho} \\
&= \frac{\boldsymbol{\gamma}\{(-\boldsymbol{Q} - \lambda \boldsymbol{I})\}^{-1}(-\boldsymbol{Q}) \exp[(\boldsymbol{Q} - \lambda \boldsymbol{I})x]\{(-\boldsymbol{Q} - \lambda \boldsymbol{I})\}\boldsymbol{e}}{1-\rho} \\
&= \frac{(\boldsymbol{\gamma} - \boldsymbol{\pi}_*) \exp[(\boldsymbol{Q} - \lambda \boldsymbol{I})x]\{(-\boldsymbol{Q} - \lambda \boldsymbol{I})\}\boldsymbol{e}}{1-\rho},
\end{aligned} \tag{A.63}$$

where the second equality ($\Pr(D < G) = 1 - \rho$) follows because (i) $\Pr(D < G)$ equals the stationary probability p_0 that an arriving customer finds the system empty and (ii) due to the Poisson-arrivals-see-time-averages (PASTA) property, p_0 in the M/PH/1 queue equals the time-average probability that the system is empty, which is given by $1 - \rho$. Also, the third equality follows because $\boldsymbol{Q}$ and $\boldsymbol{Q} - \lambda\boldsymbol{I}$ commute and the last equality ($\boldsymbol{\gamma} - \boldsymbol{\pi}_* = \boldsymbol{\gamma}\{-(\boldsymbol{Q} - \lambda\boldsymbol{I})\}^{-1}(-\boldsymbol{Q})$) is verified, noting that $\boldsymbol{\pi}_*$ satisfies (cf. (1.40) and (1.45))

$$\boldsymbol{\pi}_* = \lambda\boldsymbol{\gamma}\{-(\boldsymbol{Q} - \lambda\boldsymbol{I})\}^{-1}.$$

We then readily obtain (1.47) from (A.62) and (A.63). Also, (1.48) is obtained by integrating (1.47) and using

$$\rho \boldsymbol{\gamma}_0 \boldsymbol{e}_0 + \boldsymbol{\gamma}_1 \boldsymbol{e}_1 = \rho + (1 - \boldsymbol{\pi}_* \boldsymbol{e}) = 1,$$

where the last equality follows from (1.44).

We next consider (ii). Recall that (1.36) has been derived using the expression (1.35) for the peak AoI instead of (1.34). For the case of deterministic intergeneration times, however, it is simpler to directly work on (1.34) because it reduces to $A_{\text{peak},n} = D_n + \tau$, so that we have

$$F_{A_{\text{peak}}}(x) = \begin{cases} 0 & 0 \le x < \tau, \\ F_D(x - \tau), & x \ge \tau. \end{cases}$$

It then follows from Theorem 1.4,

$$\begin{aligned} f_A(x) &= \begin{cases} \dfrac{F_D(x)}{\tau}, & 0 \le x < \tau, \\ \dfrac{F_D(x) - F_D(x - \tau)}{\tau}, & x \ge \tau. \end{cases} \\ &= \begin{cases} \dfrac{1 - \boldsymbol{\gamma} \exp[\boldsymbol{Q}x]\boldsymbol{e}}{\tau}, & 0 \le x < \tau, \\ \dfrac{\boldsymbol{\gamma} \exp[\boldsymbol{Q}(x - \tau)]\boldsymbol{e} - \boldsymbol{\gamma} \exp[\boldsymbol{Q}\tau] \exp[\boldsymbol{Q}(x - \tau)]\boldsymbol{e}}{\tau}, & x \ge \tau, \end{cases} \end{aligned}$$

which implies (1.49). Also, (1.50) is readily obtained by integrating (1.49).

References

Asmussen, S. (1992), "Phase-type representations in random walk and queueing problems," *Ann. Prob.* **20**, 772–789.

Asmussen, S. (2003), *Applied Probability and Queues*, 2nd ed., Springer.

Bhat, U. N., & Rao, S. S. (1987), "Statistical analysis of queueing systems," *Queueing Systems* **1**(3), 217–247.

Brill, P. H., & Posner, M. J. M. (1977), "Level crossings in point processes applied to queues: Single-server case," *Oper. Res.* **25**, 662–674.

Champati, J. P., Al-Zubaidy, H., & Gross, J. (2019), On the distribution of AoI for the GI/GI/1/1 and GI/GI/1/2* systems: Exact expressions and bounds, *in* "Proc. of IEEE INFOCOM 2019," pp. 37–45.

Cohen, J. W. (1977), "On up- and downcrossing," *J. Appl. Probab.* **14**, 405–410.

Costa, M., Codreanu, M., and Ephremides, A. (2016), "On the age of information in status update systems with packet management," *IEEE Trans. Inf. Theory* **62**(4), 1897–1910.

Costa, M., Valentin, S., & Ephremides, A. (2015), On the age of channel state information for a finite-state Markov model, *in* "Proc. of IEEE ICC 2015," pp. 4101–4106.

El-Taha, M., & Stidham Jr., S. (1999), *Sample-Path Analysis of Queueing Systems*, Kluwer Academic Publishers.

Inoue, Y. (2018), Analysis of the age of information with packet deadline and infinite buffer capacity, *in* "Proc. of IEEE ISIT 2018," pp. 2639–2643.

Inoue, Y. (2020), The probability distribution of the AoI in queues with infinitely many servers, *in* "Proc. of The 3rd Age of Information Workshop (AoI'20)," pp. 297–302.

Inoue, Y., Masuyama, H., Takine, T., & Tanaka, T. (2019), "A general formula for the stationary distribution of the age of information and its application to single-server queues," *IEEE Trans. Inf. Theory* **65**(12), 8305–8324.

Inoue, Y., & Takine, T., (2019), AoI perspective on the accuracy of monitoring systems for continuous-time Markovian sources, *in* "Proc. of The 2nd Age of Information Workshop (AoI '19)."

Jiang, Y., & Miyoshi, N. (2020), "Joint performance analysis of ages of information in a multi-source pushout server," *arXiv preprint arXiv:2006.03404*.

Kaul, S. K., Yates, R. D., & Gruteser, M. (2012a), Status updates through queues, *in* "Proc. of IEEE CISS 2012."

Kaul, S., Yates, R., & Gruteser, M. (2012b), Real-time status: How often should one update?, *in* "Proc. of IEEE INFOCOM 2012," pp. 2731–2735.

Kleinrock, L. (1975), *Queueing Systems, Volume I: Theory*, John Wiley & Sons.

Kosta, A., Pappas, N., Ephremides, A., & Angelakis, V. (2020a), "The cost of delay in status updates and their value: Non-linear ageing," *IEEE Trans. Commun.* **68**(8), 4905–4918.

Kosta, A., Pappas, N., Ephremides, A., & Angelakis, V. (2020b), Non-linear age of information in a discrete time queue: Stationary distribution and average performance analysis, *in* "Proc. of ICC 2020."

Sun, Y., & Cyr, B. (2018), Information aging through queues: A mutual information perspective, *in* "Proc. of SPAWC 2018."

Sun, Y., Polyanskiy, Y., & Uysal, E. (2020), "Sampling of the Wiener process for remote estimation over a channel with random delay," *IEEE Trans. Inf. Theory* **66**(2), 1118–1135.

Sun, Y., Uysal-Biyikoglu, E., Yates, R., Koksal, C. M., & Shroff, N. B. (2017), "Update or wait: How to keep your data fresh," *IEEE Trans. Inf. Theory* **63**(11), 7492–7508.