Age of Information: A New Concept, Metric, and Tool

Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis

Kosta, A., Pappas, N., Angelakis, V., (2017), Age of Information: A New Concept, Metric, and Tool, *Foundations and Trends in Networking*, 12(3), 162-259. https://doi.org/10.1561/1300000060

Original publication available at:
https://doi.org/10.1561/1300000060

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

Age of Information: A New Concept, Metric, and Tool

Antzela Kosta
Dept. of Science and Technology
Linköping University, Sweden
antzela.kosta@liu.se

Nikolaos Pappas
Dept. of Science and Technology
Linköping University, Sweden
nikolaos.pappas@liu.se

Vangelis Angelakis
Dept. of Science and Technology
Linköping University, Sweden
vangelis.angelakis@liu.se

November 28, 2017

Abstract

Age of information (AoI) was introduced in the early 2010s as a notion to characterize the freshness of the knowledge a system has about a process observed remotely. AoI was shown to be a fundamentally novel metric of timeliness, significantly different, to existing ones such as delay and latency. The importance of such a tool is paramount, especially in contexts other than transport of information, since communication takes place also to control, or to compute, or to infer, and not just to reproduce messages of a source. This volume comes to present and discuss the first body of works on AoI and discuss future directions that could yield more challenging and interesting research.

1 Introduction

The concept of *Age of Information* (AoI) was introduced in 2011 in [31] to quantify the freshness of the knowledge we have about the status of a remote system. More specifically, AoI is the time elapsed since the generation of the last successfully received message containing update information about its source system. Utilizing a simple communication system model, in a series of papers ([32], [33], [30], and [58]), the first group of characterizations of the Age of Information metric had appeared by 2012. Since then, AoI has attracted a vivid interest, with over 50 publications, in the last six years [1].

The attention AoI has been receiving is due to two factors. The first is the sheer novelty brought by AoI in characterizing the freshness of information versus for example that of the metrics of delay or latency. Second, the need and importance of characterizing the

[1] In this volume we take into consideration works that have been published no later than June 2017.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

freshness of such information is paramount in a wide range of information, communication, and control systems. By now, age has been studied with considerable diversity of systems, being as a concept, a performance metric, and a tool.

The purpose of this volume is to present a critical summary of this first body of works performed on AoI and discuss future research directions. Already at this early point we need to put down our first disclaimer: we have chosen to treat the early works with significantly more detail, going deeper in the derivations and presenting more results and insights from them than we do with more recent works. The reason for this is to achieve a tutorial nature in the volume, which can provide a solid ground of the AoI as a concept. Moreover, the first works, which we chose to present in more detail than the rest, aim to provide fundamentally new knowledge in the premise of maintaining information fresh in a system. This basic goal opens up a wide range of communication contexts that span from estimation and prediction, to applications such as vehicular networks and information caching, to name a few.

With this in mind, we begin this volume presenting the AoI **concept** as it was originally introduced. For this, we discuss the original models of Kaul, Gruteser and Yates of [33], considering a system where a source is transmitting packets containing status updates to a destination. The analysis presented is based on a simple queueing model. Already in that work the minimization of AoI was shown to be non-trivial for the source sampling methods studied. However, it had already become clear that timely updating a destination about a remote system is neither the same as maximizing the utilization of the communication system, nor of ensuring that generated status updates are received with minimum delay. This is because utilization can be maximized by making the source send updates as fast as possible which would lead to the destination receiving delayed statuses because messages are backlogged in the communication system studied. In this case, delay suffered by the stream of status updates can be reduced by decreasing the rate of updates. Alternatively, decreasing the update rate can also lead to the destination having unnecessarily outdated status information because of lack of updates.

AoI has spawned relevant performance metrics that are more tractable such as the Peak Age of Information or the Cost of Update Delay, opening even more research opportunities. Under the timely update context, the relevant timeliness **performance metrics** should be kept at values that ensure high freshness of information. Already the first AoI lower bound had appeared in [33] and we discuss it in Section 2. We then continue the discussion on the early works of AoI as a performance metric in Section 3, where we present the case of AoI for multiple sources, its use in scheduling, and demonstrate packet management techniques that have been employed. Section 4 treats AoI as a metric for rate control, addresses the case of packets with deadlines, and presents an optimal policy for optimizing age, throughput, and delay.

Keeping the AoI metrics low is of high interest when AoI is being treated as a **tool** to facilitate the timely update of information that will eventually improve performance

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

metrics in different contexts. Consider for example remote estimation; if the process under observation consists of highly correlated data, then the frequency of generation and transmission of updates can be significantly reduced without affecting the timeliness of the information at a remote receiver. In Section 5, we discuss three domains in which AoI has been treated as a tool: Channel State Information (CSI) estimation, energy harvesting, and scheduling.

Recent works that have appeared in the time of writing of this volume have been categorized and treated in Section 6. Finally, in Section 7, we provide a brief discussion of indicative future topics on which the AoI can contribute. The topics we cover there are not a complete list, as there is an immense wealth of possible problems associated with the notion of timeliness as captured by age, in the form of either a tool, a performance metric, or even a concept.

2 The Introduction of the Concept

With the introduction of Age of Information (AoI) in [31] and the subsequent works of Kaul, Yates, and Gruteser, it became clear that this new concept is novel compared to, at that time, existing measures of information timeliness. Based on the seminal work of [33], in this section we give to the reader the fundamentals of AoI for a status update system under the queueing models assumed therein. As we already noted in the introduction, we provide a good level of detail, for this first work. This is to give solid insight and background to the reader and also to reflect the value of the contribution to the topic of information timeliness.

2.1 Age of Information

Consider a system comprising two nodes. A stochastic process $X(t)$ is observed by the source node, that extracts samples. These are assumed to carry information about the status of the process at the source node. Assuming this status information is needed at the other node, each collected sample needs to be transmitted over a communication link to that destination. At the transmitter of the source node there is a buffer which stores the samples in the form of packets containing (i) the value of the process $X(t_i)$, at time t_i when the ith sample was extracted and (ii) the timestamp t_i. Packets are being sent along the communication link of the two nodes, which is assumed error-free. Each such *packet* arriving at the destination, is said to provide a *status update* and these two terms are used interchangeably.

A simple queueing model is employed, as shown in Figure 2.1, where all packets $i = 1, 2, \ldots$ generated at the source s need to reach the destination denoted by d. The storage of the packets at the queue is instantaneous, thus the packet arrivals at the queue

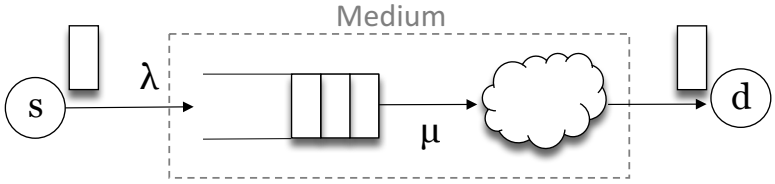

Figure 2.1: The basic system model.

are characterized by the sampling rate of $X(t)$ and so the terms *status update generation* and *packet arrival* can be used interchangeably. Consider that the status update generation is modeled as a stochastic process of average rate λ and packets are then transmitted with an average service rate μ. Later in this section, we will discuss cases that the arrival, or service is a deterministic process.

The freshness of the knowledge the destination has about the status of the source node is captured by the concept of the AoI. With AoI this freshness is quantified, at any moment, as the time that elapsed since the last received status update was generated by the source.

Definition 2.1.1 (Age of Information – AoI)**.** Consider a system comprising a source-destination communication pair. Let $t_k^{'}$ be the times at which the status updates are received at the destination. At time ξ, the index of the most recently received update is

$$N(\xi) = \max\{k | t_k^{'} \leq \xi\}, \tag{2.1}$$

and the timestamp of the most recently received update is

$$u(\xi) = t_{N(\xi)}. \tag{2.2}$$

The *Age of Information* (AoI) of the source s at destination d is then defined as the random process

$$\Delta(t) = t - u(t). \tag{2.3}$$

Notice that the terms *Age of Information*, *status age*, or plain *age*, are used interchangeably throughout the remaining of this volume.

Figure 2.2 shows an illustrative example of the evolution of AoI in time. Without loss of generality, assume that at $t = 0$ we start observing the system, the queue is empty, and the AoI at the destination is $\Delta(0) = \Delta_0$. Status update i is generated at time t_i and is received by the destination at time $t_i^{'}$. Between $t_{i-1}^{'}$ and $t_i^{'}$, where there is an absence of updates at the destination, the AoI increases linearly with time. Upon reception of a status update the AoI is reset to the delay that the packet experienced going through the transmission system.

The ith interarrival time is defined as the time elapsed between the generation of update i and the previous update generation, thus Y_i is the random variable

$$Y_i = t_i - t_{i-1}. \tag{2.4}$$

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

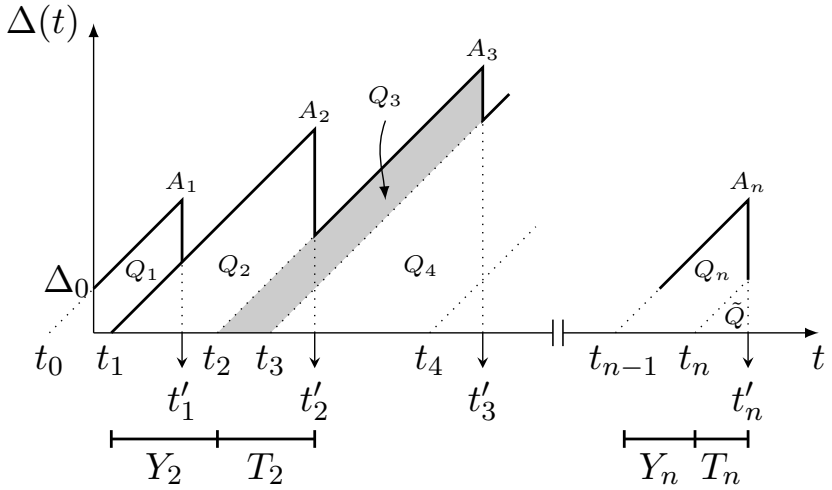

Figure 2.2: Example of age evolution.

Moreover,

$$T_i = t'_i - t_i, \tag{2.5}$$

is the system time of update i, corresponding to the sum of the queue waiting time and the service time. Assuming that the observation interval is from $t = 0$ to $t = \mathcal{T} = t'_n$, we denote

$$N(\mathcal{T}) = \max\{n | t_n \leq \mathcal{T}\}, \tag{2.6}$$

the number of arrivals by time $\mathcal{T}$. Then, at times t'_i for $i = \{1, 2, \ldots, N(\mathcal{T})\}$ the age $\Delta(t'_i)$ is reset to $T_i = t'_i - t_i$. The reduction in age with each received update captures the freshness of the information of the source status at the destination. For any time that we do not have an update received at the destination, i.e. times that do not belong to the set $\mathcal{I} \doteq \{t'_1, t'_2, \ldots t'_{N(\mathcal{T})}\}$ the age increases as time passes by. Therefore, the age process exhibits the sawtooth pattern as shown in Figure 2.2.

Analysis of the average Age of Information

In the context we have been discussing so far, the objective of the communication system would be to maintain the information from the source as fresh as possible. Ensuring the average AoI of such a system is small corresponds to maintaining information about the status of the source at the destination fresh. In what follows, we will go into the first characterizations for different queueing disciplines and in subsequent sections, treating AoI as a metric, we will see mechanisms that have been developed to keep the average AoI low.

Given an age process $\Delta(t)$ and assuming ergodicity, the average age can be calculated using a sample average that converges to its corresponding stochastic average.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

Definition 2.1.2 (Time average Age of Information)**.** For an interval of observation $(0, \mathcal{T})$, the time average age of a status update system is

$$\Delta_{\mathcal{T}} = \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \Delta(t)\, dt. \tag{2.7}$$

The integral in (2.7) can be calculated as the area under $\Delta(t)$. Then, the time average age can be rewritten as a sum of disjoint geometric parts. Starting from $t = 0$, the area is decomposed into the polygon area Q_1, the trapezoids Q_i for $i = 2, 3, \ldots N(\mathcal{T})$, and the triangular area of width T_n that we denote $\tilde{Q}$. Then, the decomposition of $\Delta_{\mathcal{T}}$ yields

$$\begin{aligned}\Delta_{\mathcal{T}} =& \frac{1}{\mathcal{T}} \left(Q_1 + \tilde{Q} + \sum_{i=2}^{N(\mathcal{T})} Q_i \right) \\ =& \frac{Q_1 + \tilde{Q}}{\mathcal{T}} + \frac{N(\mathcal{T}) - 1}{\mathcal{T}} \frac{1}{N(\mathcal{T}) - 1} \sum_{i=2}^{N(\mathcal{T})} Q_i.\end{aligned} \tag{2.8}$$

The time average $\Delta_{\mathcal{T}}$ tends to the ensemble *average age* as $\mathcal{T} \to \infty$, i.e.,

$$\Delta = \lim_{\mathcal{T} \to \infty} \Delta_{\mathcal{T}}. \tag{2.9}$$

Note that the term $(Q_1 + \tilde{Q})/\mathcal{T}$ goes to zero as $\mathcal{T}$ grows and also let

$$\lambda = \lim_{\mathcal{T} \to \infty} \frac{N(\mathcal{T})}{\mathcal{T}} \tag{2.10}$$

be the steady state rate of status updates generation. Furthermore, using the definitions (2.4) and (2.5) of the interarrival and system times respectively, we can write the trapezoidal areas as

$$Q_i = \frac{1}{2}(T_i + Y_i)^2 - \frac{1}{2} T_i^2 = Y_i T_i + Y_i^2/2. \tag{2.11}$$

Then, substituting (2.8), (2.10), and (2.11) to (2.9) the *average Age of Information* in the status update system of Figure 2.1 is given by

$$\Delta = \frac{\mathbb{E}[Q]}{\mathbb{E}[Y]} = \frac{\mathbb{E}[YT] + \mathbb{E}[Y^2]/2}{\mathbb{E}[Y]}, \tag{2.12}$$

where $\lambda = 1/\mathbb{E}[Y]$ and $\mathbb{E}[\cdot]$ is the expectation operator. Notice that ergodicity has been assumed for the stochastic process $\Delta(t)$ but no assumptions regarding the distribution of the random variables Y and T, have been made nor any specific service policy has been considered. This result also holds when the system is shared among multiple traffic streams.

Observe that the random variables Y (interarrival time) and T (system time) are dependent and this complicates the calculations of the average age in the general case, since we do not know their joint distribution. Intuitively, for a fixed service rate, reducing

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

interarrival times correspond to packets filling up the system. This increased traffic leads to larger system times. On the other hand, larger interarrival times allow the queue to empty and consequently the delays are smaller. Thus, Y and T are negatively correlated. In the next section we will discuss cases where Y and T are conditionally independent.

2.2 First Queue-theoretic System Abstractions

In the first AoI works, the communication model considered was a simple queueing system, since queueing theory has been a core methodological framework for analysing delay. In this framework, a first step is to identify the nature of the arrival process, the probability distribution of the service times, the number of servers, and the queue discipline. In [33] three simple models were studied, the M/M/1, the M/D/1, and the D/M/1, under the first-come-first-served (FCFS) discipline. Here, we illustrate the analyses of these cases and later, in subsequent sections, we present more complicated models that can capture more realistically the medium through which the signal is transmitted to reach the destination, especially in the wireless case.

The M/M/1 system model

Consider an M/M/1 system where packets are served with an FCFS policy. Such a model captures a system with limited resources that consists of one source and one server, where status updates are generated according to a Poisson process with mean arrival rate λ. Thus, the interarrival times Y are independent and identically distributed (i.i.d.) exponential random variables with $\mathbb{E}[Y] = 1/\lambda$. Furthermore, the service times are i.i.d. exponentials with mean $1/\mu$ and the server utilization is defined as $\rho = \frac{\lambda}{\mu}$.

The average age was given in (2.12), so the terms $\mathbb{E}[Y^2]$ and $\mathbb{E}[YT]$ need to be calculated. Since Y is exponentially distributed with mean arrival rate λ, we have $\mathbb{E}[Y^2] = 2/\lambda^2$. For $\mathbb{E}[YT]$, consider that the system time of update i is

$$T_i = W_i + S_i, \tag{2.13}$$

where W_i is the waiting time and S_i is the service time of update i. Since, the service time S_i is independent of the ith interarrival time Y_i, we can write

$$\mathbb{E}[T_i Y_i] = \mathbb{E}[(W_i + S_i)Y_i] = \mathbb{E}[W_i Y_i] + \mathbb{E}[S_i]\mathbb{E}[Y_i], \tag{2.14}$$

where $\mathbb{E}[S_i] = 1/\mu$ and $\mathbb{E}[Y_i] = 1/\lambda$. Moreover, we can express the waiting time of update i as the remaining system time of the previous update minus the elapsed time between the generation of updates $(i-1)$ and i, i.e.,

$$W_i = (T_{i-1} - Y_i)^+. \tag{2.15}$$

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

Note that if the queue is empty then $W_i = 0$. Also note that when the system reaches steady state the system times are stochastically identical, i.e., $T =^{st} T_{i-1} =^{st} T_i$. Additionally, the probability density function (pdf) of the system time T for the M/M/1 is [44]

$$f_T(t) = \mu(1-\rho)e^{-\mu(1-\rho)t}, \quad t \geq 0. \tag{2.16}$$

Thus, the conditional expectation of the waiting time W_i given $Y_i = y$ can be obtained as

$$\begin{aligned}\mathbb{E}[W_i|Y_i = y] &= \mathbb{E}[(T_{i-1} - y)^+|Y_i = y] = \mathbb{E}[(T-y)^+] \\ &= \int_y^\infty (t-y) f_T(t)\, \mathrm{d}t = \frac{e^{-\mu(1-\rho)y}}{\mu(1-\rho)}.\end{aligned} \tag{2.17}$$

The expectation $\mathbb{E}[W_i Y_i]$ is then obtained as

$$\mathbb{E}[W_i Y_i] = \int_0^\infty y\, \mathbb{E}[W_i|Y_i = y] f_{Y_i}(y)\, \mathrm{d}y = \frac{\rho}{\mu^2(1-\rho)}. \tag{2.18}$$

From (2.18), (2.14) and (2.12), the average AoI is obtained as

$$\Delta_{\mathrm{M/M/1}} = \frac{1}{\mu}\left(1 + \frac{1}{\rho} + \frac{\rho^2}{1-\rho}\right). \tag{2.19}$$

We are interested in minimizing the average age $\Delta_{\mathrm{M/M/1}}$ with respect to the server utilization ρ. Assuming fixed average service rate μ, the optimal server utilization corresponds to an optimal arrival rate λ. In this case, we assume that we are able to generate status updates as frequently as we want by extracting instances of the random process $X(t)$ under observation. From (2.19), we can obtain that the optimal server utilization is $\rho^* \approx 0.53$. At the optimal server utilization the average number of packets in the system is $\rho^*/(1-\rho^*) \approx 1.13$.

We observe that optimizing the timeliness of status updates through the AoI results in a policy that might sound controversial until this point. The minimum age is achieved by keeping the server idle $\approx 47\%$ of the time, which differs from a policy that sends updates as fast as possible ($\rho \to 1$) in order to maximize throughput or follows a conservative approach with ρ close to 0 to minimize delay. With the optimal arrival rate the server is being busy slightly more than being idle.

The M/D/1 system model

In the M/D/1 model status updates are generated according to a Poisson process with average rate λ and the service time is deterministic, which in (2.13) means that $S_i = D$ for all updates i with D being a fixed value. Pursuing the objective to characterize and then minimize the average age (2.12) the term $\mathbb{E}[YT]$ is needed, as in the M/M/1 case. Consider the system time of update i to be $T_i = W_i + D$. Then,

$$\mathbb{E}[T_i Y_i] = \mathbb{E}[W_i Y_i] + D\, \mathbb{E}[Y_i], \tag{2.20}$$

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

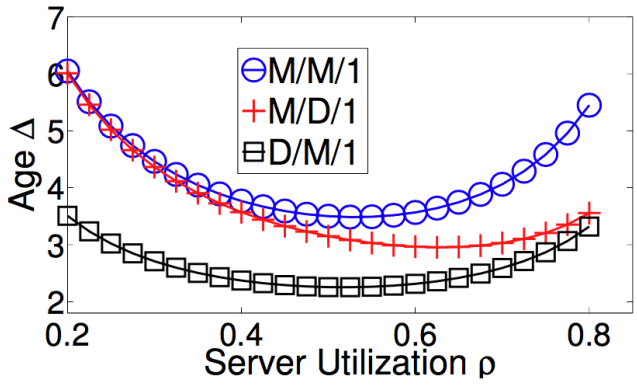

Figure 2.3: Average age vs server utilization for the M/M/1, M/D/1, and D/M/1 systems and fixed service rate $\mu = 1$ [33].

where $\mathbb{E}[Y_i] = 1/\lambda$. Similar to (2.17), we can write $\mathbb{E}[W_i|Y_i = y] = \mathbb{E}[(T-y)^+] = \mathbb{E}[(W + D - y)^+]$ and compute this term using the expectation of the waiting time W for the M/D/1 queue given by [**Table 7.1**, 42]

$$\mathbb{E}[W] = \frac{\mathbb{E}[S]\rho}{2(1-\rho)}. \tag{2.21}$$

Having $\mathbb{E}[W_i|Y_i = y]$ we can use the iterated expectation to derive $\mathbb{E}[W_i Y_i]$. Due to the complexity of the analysis the exact AoI expression is not provided in the literature.

The server utilization that minimizes the average age was numerically evaluated to be $\rho^* \approx 0.625$ in [33]. There, the authors illustrated the change of average age at the destination as a function of ρ for service rate $\mu = 1$ (see Figure 2.3). Observe that the performance of the M/M/1 and M/D/1 systems is similar for small values of the arrival rate λ, where the waiting time is less sensitive to the service rate. As ρ becomes larger the gap between the M/M/1 and the M/D/1 system age increases. For these ρ values, we have more packets waiting for service and the deterministic server of the M/D/1 queue performs better with backlog phenomena. Overall, the M/D/1 system achieves smaller average age than the M/M/1 for all ρ.

The D/M/1 system model

In a D/M/1 model the status updates are generated at a deterministic period D and service times are exponentially distributed with mean $1/\mu$. We have that $\mathbb{E}[Y] = D$, $\mathbb{E}[Y^2]/2 = D^2/2$, and $\mathbb{E}[YT] = D\mathbb{E}[T]$, so the only unknown term for the calculation of the average age is the expectation of the system time T.

Consider the system time of update i to be $T = W + S$. The average system time can

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

be written as

$$\mathbb{E}[T] = \mathbb{E}[W] + \mathbb{E}[S] = \frac{\beta}{\mu(1-\beta)} + \frac{1}{\mu}, \tag{2.22}$$

where $0 \leq \beta \leq 1$ is the solution to the equation $\beta = L_X(\mu(1-\beta))$, with $L_X(\cdot)$ being the Laplace transform of the distribution of the interarrival times [42], expanded in detail in [33].

As a result, the average AoI is calculated as

$$\Delta_{\mathrm{D/M/1}} = \frac{1}{D}\left[\frac{D^2}{2} + D\,\mathbb{E}[T]\right] = \frac{1}{\mu}\left[\frac{1}{2\rho} + \frac{1}{1-\beta}\right]. \tag{2.23}$$

The change of age with the server utilization ρ for D/M/1 is illustrated in Figure 2.3. The optimal server utilization is $\rho^* = 0.515$. The utilization that minimizes the age for the D/M/1 queue is similar to the optimal ρ for the M/M/1 queue, but the D/M/1 system leads to almost 50% decrease of age for all ρ.

Just-in-time lower bound

To derive a lower bound of age for the systems we presented earlier, consider that the source can observe the state of the queue and submit a new update as soon as the previous update completes service. Thus, there is no queueing of updates in the system and the server is always busy. Since each delivered status update is as fresh as possible, the average AoI obtained for this system is a lower bound to the age for any queue in which updates are generated as a stochastic process independent of the current state of the queue.

To characterize the described system mathematically, we have that $t_i = t'_{i-1}$ for all packets i. For any packet i, $W_i = 0$, and also $T_i = S_i$, $\lambda = 1/\mathbb{E}[S]$, and $Y_{i+1} = S_i$. Then, the average age in (2.12) can be written as

$$\Delta = \frac{(\mathbb{E}[S])^2 + \mathbb{E}[S^2]/2}{\mathbb{E}[S]}. \tag{2.24}$$

For a system with memoryless service at rate $\mu = 1/\mathbb{E}[S]$, we get $\Delta = 2/\mu$.

The property $\mathbb{E}[S^2] \geq (\mathbb{E}[S])^2$ combined with (2.24) yields the lower bound

$$\Delta_{\text{just-in-time LB}} = \frac{3\mathbb{E}[S]}{2} = \frac{3}{2\mu}. \tag{2.25}$$

In Section 4.1, we will present under which conditions the just-in-time policy is optimal and provide alternative policies that achieve better performance.

2.3 Summary

Having presented the very first model on which the notion of Age of Information was developed, and discussed three basic queueing models, in the following we open up our

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

discussion to different models that have enabled the study AoI and its application in a much wider range of application domains. Indeed in what follows, although we do not yet depart from the queueing models, we present works with multiple sources, multiple servers, begin the discussion on AoI in scheduling policies and introduce the first by-product metric, that of Peak Age of Information, which enables more tractable analysis.

3 The Early Works

The first works [31, 32, 33, 30, 58] on Age of Information can be considered to provide a new framework in the context of timely communicating information. The works [30, 33, 58] took a queue based theoretical viewpoint of the system model. A number of subsequent works departed from this model, however there remains a strand of research working on adaptations of the original model of Figure 2.1. In the previous section, we illustrated first results for different queueing models of the FCFS discipline. It remains of interest to investigate AoI optimality under different queue disciplines and characteristics. This section, aims to address these aspects.

We start the presentation with a system of multiple independent sources that coexist in the network. As a first step the multiple sources are served under an FCFS policy. We then move to different policies, that can improve the AoI performance, considering the following three cases. The first, is based on the last-come-first-served (LCFS) queue discipline with and without preemption. The second, considers systems with different availabilities of resources (servers), taking also into consideration path diversity. The third, focuses on queues of finite size and in this case, we discuss packet management schemes of the literature. Finally, a new performance metric called *Peak Age of Information* introduced in [11] is presented.

3.1 Sharing the System Among Multiple Traffic Streams

Here we consider a system with multiple independent sources providing status updates to a common destination through an FCFS-served M/M/1 queue, as studied in [58]. The analysis provided in Subsection 2.1 holds even when the system is shared among multiple traffic streams. Therefore, equation (2.12) can be used to derive the average AoI for each independent source.

We consider an M/M/1 system where status updates are generated according to a Poisson process with mean rate λ_i for source i, thus, we have $\mathbb{E}[Y] = 1/\lambda_i$ and $\mathbb{E}[Y^2] = 2/\lambda_i^2$. Furthermore, the service times are i.i.d. exponentially distributed with mean $\mathbb{E}[S] = 1/\mu$ and the server utilization of source i is defined as $\rho_i = \lambda_i/\mu$. The overall load is $\rho = \sum_{i=1}^{N} \rho_i$.

The difficulty of the analysis lies in calculating the expectation $\mathbb{E}[YT]$, which is equal to $\mathbb{E}[YW] + \mathbb{E}[S]\mathbb{E}[Y]$. Let Y_j, W_j and T_j be the interarrival time, waiting time, and system

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

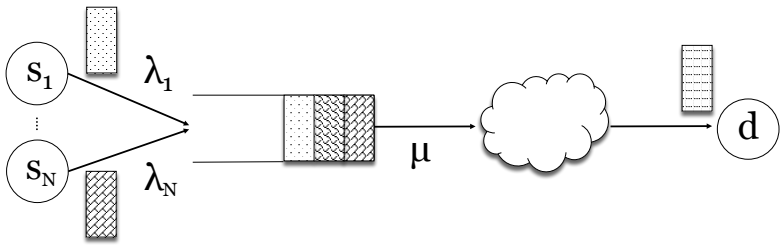

Figure 3.1: System model with multiple independent sources.

time, respectively, of the jth packet of source i[2]. We proceed with the characterization of W_j via the partition

$$B_j = \{Y_j < T_{j-1}\}, \quad L_j = \{T_{j-1} < Y_j\}, \tag{3.1}$$

where L_j is the complementary event of B_j. Then, we have

$$\mathbb{E}[Y_j W_j] = \mathbb{E}[Y_j W_j | B_j]\mathrm{P}[B_j] + \mathbb{E}[Y_j W_j | L_j]\mathrm{P}[L_j]. \tag{3.2}$$

To continue with the calculations consider the following properties of Poisson processes and exponential random variables [44].

Lemma 3.1.1. Let X_1 and X_2 be independent exponential random variables with $\mathbb{E}[X_i] = 1/\alpha_i$. Let $V = X_2 - X_1$.

(a) $\mathrm{P}[X_1 < X_2] = \alpha_1/(\alpha_1 + \alpha_2)$.

(b) Given $X_1 < X_2$, X_1 and V are conditionally independent and have conditional exponential probability density functions (pdfs)

$$f_{X_1|X_1<X_2}(x) = (\alpha_1 + \alpha_2)e^{-(\alpha_1+\alpha_2)x}, \qquad x \geq 0,$$
$$f_{V|X_1<X_2}(v) = \alpha_2 e^{-\alpha_2 v}, \qquad v \geq 0.$$

Lemma 3.1.2. Given a Poisson process $N(t)$ with average rate λ and an independent exponential random variable X with parameter α, the number of arrivals $N(X)$ in the interval $[0, X]$ has the geometric pmf

$$\mathrm{P}_{N(X)}(n) = (1-\gamma)\gamma^n, \qquad n \geq 0,$$

with $\gamma = \lambda/(\alpha + \lambda)$.

[2]For presentation clarity, we drop the index denoting the source and focus on the packet index.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

The system time T_{j-1} depends only on packets that are generated prior to the $(j-1)$th packet, therefore it is independent of Y_j. Using this independence and Lemma 3.1, we have $\mathrm{P}[B_j] = \rho_i/(1-\rho_{-i})$, where ρ_{-i} is defined as the aggregate other-source load

$$\rho_{-i} = \rho - \rho_i = \sum_{l \neq i} \rho_l. \tag{3.3}$$

Given the event B_j, assume that packet $(j-1)$ is still in the system when packet j is generated by the same source i. Then, the waiting time W_j is the remaining system time of packet $(j-1)$ plus the system time of packets from other sources that arrive during the interarrival period Y_j of source i. More specifically, $M = N_{-i}(Y_j)$ denotes the number of packets of sources other than i arriving during the interarrival period Y_j; $S_1, S_2, \ldots, S_M$ denote the service times of those packets. Then,

$$W_j = (T_{j-1} - Y_j) + \sum_{k=1}^{M} S_k. \tag{3.4}$$

It follows that $\mathbb{E}[Y_j W_j | B_j] = E_1 + E_2$ where

$$E_1 = \mathbb{E}[Y_j (T_{j-1} - Y_j) | B_j], \tag{3.5a}$$

$$E_2 = \mathbb{E}\left[Y_j \sum_{k=1}^{M} S_k | B_j \right]. \tag{3.5b}$$

Using Lemma 3.1(b) for the first term we obtain

$$\begin{aligned} E_1 &= \mathbb{E}[(T_{j-1} - Y_j)|B_j]\mathbb{E}[Y_j|B_j] \\ &= \frac{1}{\mu - \lambda}\left(\frac{1}{\lambda_i + (\mu - \lambda)}\right) = \frac{1}{\mu^2(1-\rho)(1-\rho_{-i})}. \end{aligned} \tag{3.6}$$

For the second term, using iterated expectation we have

$$\begin{aligned} E_2 &= \int_0^\infty \mathbb{E}\left[Y_j \sum_{k=1}^{M} S_k | B_j, Y_j = y \right] f_{Y_j|B_j}(y)\, \mathrm{d}y \\ &= \int_0^\infty \mathbb{E}\left[y \sum_{k=1}^{M} S_k | Y_j = y \right] f_{Y_j|B_j}(y)\, \mathrm{d}y. \end{aligned} \tag{3.7}$$

The conditional expectation in the integral yields

$$\begin{aligned} \mathbb{E}\left[y \sum_{k=1}^{M} S_k | Y_j = y \right] &= y\, \mathbb{E}[M|Y_j = y]\mathbb{E}[S_k|Y_j = y] = y(\lambda_{-i} y)(1/\mu) \\ &= \rho_{-i}\, y^2, \end{aligned} \tag{3.8}$$

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

where the independence of S_k and Y_j allows us to write $\mathbb{E}[S_k|Y_j = y] = \mathbb{E}[S_k] = 1/\mu$, and the conditional expectation of the number of packets M of sources other than i, arriving during the interarrival period $Y_j = y$ is $\mathbb{E}[M|Y_j = y] = \lambda_{-i}\, y$. By Lemma 3.1, $f_{Y_j|B_j}$ is exponentially distributed with average rate $\alpha = \lambda_i + (\mu - \lambda_i - \lambda_{-i}) = \mu - \lambda_{-i}$. This implies

$$E_2 = \rho_{-i} \int_0^\infty y^2 \alpha e^{-\alpha y}\, \mathrm{d}y = \frac{2\rho_{-i}}{\alpha^2} = \frac{2\rho_{-i}}{\mu^2(1-\rho_{-i})^2}. \tag{3.9}$$

We write the expectation $\mathbb{E}[Y_j W_j|B_j]$ using (3.6) and (3.9) as

$$\mathbb{E}[Y_j W_j|B_j] = \frac{1}{\mu^2}\left[\frac{2\rho_{-i}}{(1-\rho_{-i})^2} + \frac{1}{(1-\rho)(1-\rho_{-i})}\right]. \tag{3.10}$$

Given the event L_j, packet $(j-1)$ has already departed the system when packet j is generated. In this case, the waiting time of packet j depends on the number of other-source packets in the system when packet j is generated. By Lemma 3.2, the number of other-source packets in the system is geometrically distributed and independent to both the additional delay $Y_j - T_{j-1}$ until the arrival of packet j and the prior system time T_{j-1}. Then, we obtain (see details in [61])

$$\begin{aligned}\mathbb{E}[Y_j W_j|L_j] &= \mathbb{E}[(T_{j-1} + (Y_j - T_{j-1}))W_j|L_j] \\ &= \mathbb{E}[(T_{j-1} + (Y_j - T_{j-1}))|L_j]\mathbb{E}[W_j|L_j] \\ &= \left(\frac{1}{\mu - \lambda_{-i}} + \frac{1}{\lambda_i}\right)\left(\frac{\rho_{-i}}{\mu(1-\rho_{-i})}\right).\end{aligned} \tag{3.11}$$

Substituting the derived conditional expectations in (3.2) we obtain

$$\mathbb{E}[YW] = \frac{1}{\mu^2}\left[\frac{\rho_i(1-\rho\rho_{-i})}{(1-\rho)(1-\rho_{-i})^3} + \frac{\rho_{-i}}{\rho_i(1-\rho_{-i})}\right]. \tag{3.12}$$

Finally, the average AoI of source i in a shared system is given by

$$\Delta_{i,\mathrm{M/M/1}} = \frac{1}{\mu}\left[\frac{\rho_i^2(1-\rho\rho_{-i})}{(1-\rho)(1-\rho_{-i})^3} + \frac{1}{1-\rho_{-i}} + \frac{1}{\rho_i}\right]. \tag{3.13}$$

Consider a system shared by two sources. The average age of equation (3.13) is illustrated for each one of them, and for various server utilizations, as shown in Figure 3.2 [61]. The total load is fixed to $\rho = \rho_1 + \rho_2$ and the service rate is $\mu = 1$. The sum $\Delta_1 + \Delta_2$ is minimized at $\rho_1 = \rho_2 = 0.306$ yielding $\Delta_1 = \Delta_2 = 5.30$. The region of feasible age pairs (Δ_1, Δ_2) is bordered by the $\rho = 0.612$ curve and expands on the right. When a source i is limited by a load constraint $\rho_i \leq \bar{\rho}_i$ and sources other than i offer a combined load ρ_{-i}, then source i can decrease its average age Δ_i by unilaterally increasing ρ_i to $\bar{\rho}_i$. The Nash equilibrium of a system is achieved when each source i operates at its maximum allowed load $\bar{\rho}_i$ and is depicted for $\rho = 0.684$ in our case. Recall that serving only one source we

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

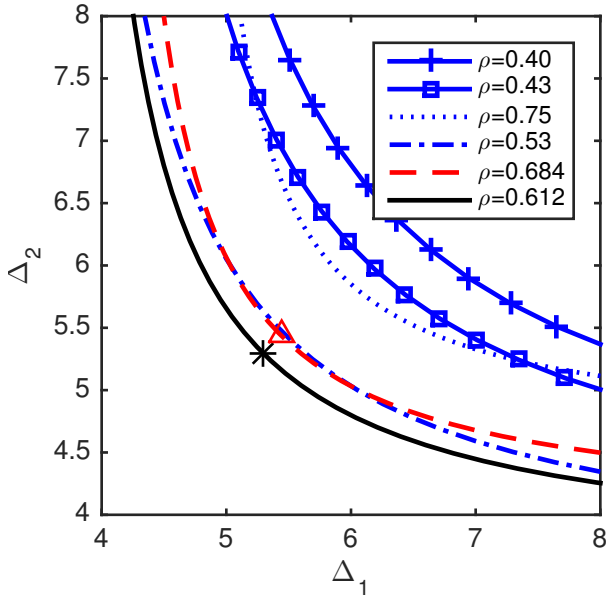

Figure 3.2: Average age for two sources sharing an M/M/1 FCFS queue with fixed service rate $\mu = 1$ and fixed total load $\rho = \rho_1 + \rho_2$.With $*$ is the minimum achievable age and with $\triangle$ the Nash equilibrium achieved by unilateral optimization [61].

obtain minimum age $\Delta_1 = 3.48$ and the optimal server utilization is $\rho = 0.53$. This implies that serving two sources through separate systems with service rate $\mu = 1/2$ each, would yield $\Delta_1 = \Delta_2 = 6.96$. We can therefore conclude that combining two sources (source 2 can be considered equivalent to an aggregate of multiple Poisson streams) in a common queue is more efficient than serving them separately.

3.2 Basic Scheduling Through Queues

An important property of the status update systems that we will consider from now on until the end of the monograph is the Markov property of the stochastic process $X(t)$ under observation. We assume that if the obtained data have this property, then only the latest status update is important and the receiver does not have to retain a history of updates. This fact agrees well with the nature of sensing and actuation applications dealing with time critical information. Therefore, if a packet arrives at the destination after a newer generated packet has already arrived, then it does not contain useful information. Thus, it is of interest to investigate the performance of the last-come-first-served (LCFS) discipline.

So far we have presented the case where a status update system is modeled as a simple M/M/1 queue and provided a closed form expression of AoI for both a single and a multiple source system. In addition, we presented the effect of the interarrival times and the service

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

times on the system performance. The work in [30] first breaks the FCFS assumption by allowing newly generated status updates to surpass older updates. The idea is to use the LCFS discipline; thus the most recent updates have the priority in transmission, and furthermore to replace queued updates with the arrival of newer ones from the same source. This packet control mechanism improves the performance of the system with respect to the Age of Information.

3.2.1 The Last-Come-First-Served Queue Discipline

Here, two LCFS systems, with and without the ability to preempt a packet in service, will be discussed. First, under LCFS without preemption, the new status packet replaces any older status packet waiting in the queue. However, it has to wait for the packet currently under service to finish. Second, under LCFS with preemption, the new packet is allowed to replace the packet currently in service.

The age process exhibits a sawtooth pattern as shown in Figure 2.2, however, in this case the interarrival times Y_i and system times T_i refer only to packets that have been served. Under these general assumptions equation (2.12) holds for the LCFS queue discipline with and without preemption. An M/M/1 system is considered, and the average age for multiple independent sources with preemption is calculated in [30].

LCFS without preemption

An example of the evolution of $\Delta(t)$ over time for the LCFS discipline without preemption is illustrated in Figure 3.3. The time instants, t_i, for $i = 1, 2, \ldots, n$ correspond to the packets i that receive service. Packets generated between time t_i and t_{i+1} may include packets from other sources that may or may not complete service. In addition, this time interval may include packets from the source under consideration that do not receive service as they are replaced by a new packet arrival. (Such updates are generated at times denoted by t_{ik} for $k = 1, 2, \ldots.$)

Using the notation of Section 2.2, under the LCFS discipline without preemption, we have that the ith interarrival time is denoted by $Y_i = t_i - t_{i-1}$. This corresponds to the time interval between the arrivals of two consecutive successfully transmitted packets. If the ith transmitted packet finds an empty system upon arrival, i.e. the $(i-1)$th transmitted packet has completed service, then Y_i is exponentially distributed with mean $1/\lambda$. On the other hand, if the ith packet is not the next to arrive after the $(i-1)$th packet, but one or more packets arrive in between them, then Y_i is a random sum of exponentially distributed interarrival times $\tilde{Y}$ with mean $1/\lambda$. Hence,

$$Y_i = \sum_{k=1}^{N_i} \tilde{Y}_k. \tag{3.14}$$

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

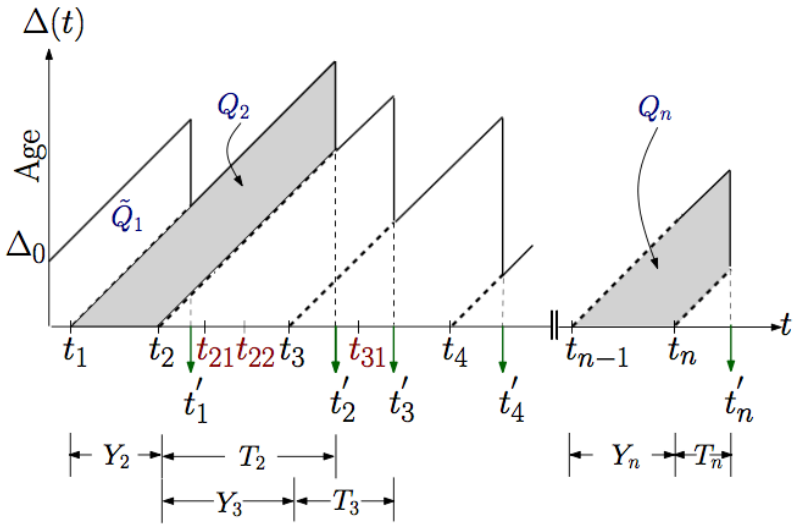

Figure 3.3: Example of age evolution of a given source for a system using LCFS *without* preemption [30].

The random variable Y_i does not have the memoryless property of the interarrival times and is difficult to characterize. This observation significantly complicates the calculations of the average age. For this reason, in Subsection 3.5 an alternative approach through a model that will be called M/M/1/2* is presented. A comprehensive analysis of an M/M/1 LCFS system with and without preemption and multiple sources can be found in [61]. In this work, the authors introduce an analytic technique called stochastic hybrid systems (SHS) [20] to determine the feasible average age region.

LCFS with preemption

Under the LCFS queue discipline with preemption, a packet arrival preempts the packet currently in service, if any. Packets arrive from one or multiple independent sources. Every packet enters the service immediately after its generation but it may or may not complete service. A new arrival and the packet being preempted may not belong to the same source.

Figure 3.4 shows an example of the evolution of $\Delta(t)$ over time for one source. The time instants, t_j, for $j = 1, 2, \ldots, n$,[3] correspond to the packets j that complete service. The interval $Y_j = t_j - t_{j-i}$ is therefore defined as the time elapsed between the generation of such packets. Let Z_j be the time interval between the departure of packet $(j-1)$ and j

$$Z_j = t'_j - t'_{j-1}. \tag{3.15}$$

Any arrivals of a given source during Z_j, other than j, are preempted. However, it is possible for arrivals of other sources to complete service. Figure 3.4 shows Z_3, which consists of a

[3]Note that the index denoting the packet is now j and the index i refers to the source.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

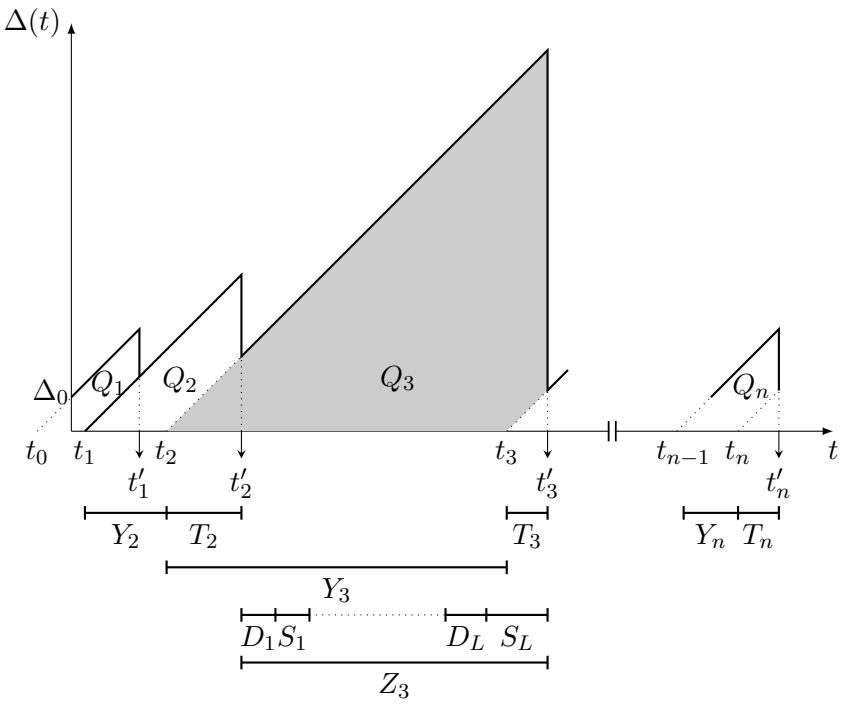

Figure 3.4: Example of age evolution of a given source for a system using LCFS *with* preemption.

random number L of time intervals. Each time interval $1 \leq k \leq L$, of length B_k, consists of an idle period of length D_k followed by a busy period of length S_k. That is

$$Z_j = \sum_{k=1}^{L} B_k = \sum_{k=1}^{L} (D_k + S_k). \tag{3.16}$$

Note that packet j arrives during S_L and then spends T_j amount of time in service.

To calculate the terms $\mathbb{E}[Y]$, $\mathbb{E}[Y^2]$ and $\mathbb{E}[TY]$ we can write

$$Y_j = T_{j-1} + Z_j - T_j. \tag{3.17}$$

Also note that when the system reaches steady state the system times are stochastically identical, i.e., $Y =^{st} Y_j$, $T =^{st} T_{j-1} =^{st} T_j$, and $Z =^{st} Z_j$. Thus,

$$\mathbb{E}[Y] = \mathbb{E}[Y_j] = \mathbb{E}[Z_j] = \mathbb{E}[Z]. \tag{3.18}$$

Since Z_j and T_j are dependent, but each one is independent of T_{j-1}, we obtain

$$\mathbb{E}[Y^2] = \mathbb{E}[Z^2] + 2\,\mathrm{Var}[T] - 2\,\mathrm{Cov}[ZT]. \tag{3.19}$$

Using the fact that Y_j and T_j are mutually independent, we have

$$\mathbb{E}[Y_j T_j] = \mathbb{E}[Y_j]\mathbb{E}[T_j] = \mathbb{E}[Y]\mathbb{E}[T] = \mathbb{E}[Z]\mathbb{E}[T]. \tag{3.20}$$

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

Finally, substituting (3.18), (3.19), and (3.20) to (2.12) the average AoI of source i for general queue models is given by

$$\Delta_{i,\text{LCFS,preempt}} = \mathbb{E}[T] + \frac{\mathbb{E}[Z^2]}{2\mathbb{E}[Z]} + \frac{\text{Var}[T] - \text{Cov}[ZT]}{\mathbb{E}[Z]}. \tag{3.21}$$

Next, we consider a specific arrival and service process, that is, an M/M/1 system where status updates are generated according to a Poisson process with mean rate λ_i for source i, as illustrated in Figure 3.1. The overall arrival rate is $\lambda = \sum_{i=1}^{N} \lambda_i$. Furthermore, the service times are i.i.d. exponentially distributed with mean $\mathbb{E}[S] = 1/\mu$ and the server utilization of source i is defined as $\rho_i = \lambda_i/\mu$. Given the general result in (3.21) we need to calculate the terms $\mathbb{E}[T]$, $\mathbb{E}[T^2]$, $\mathbb{E}[Z]$, $\mathbb{E}[Z^2]$, $\mathbb{E}[ZT]$. Using the analysis in [30], we have that T_j is an exponentially distributed random variable with

$$\mathbb{E}[T_j] = \frac{1}{\lambda+\mu}, \quad \text{and} \quad \mathbb{E}[T_j^2] = \frac{2}{(\lambda+\mu)^2}. \tag{3.22}$$

Moreover, the first and second moment of the interdeparture time of source i are given by

$$\mathbb{E}[Z_j] = \frac{\mu+\lambda}{\lambda_i\mu}, \quad \text{and} \quad \mathbb{E}[Z_j^2] = 2\frac{\lambda}{\lambda_i}\left(\frac{\lambda}{\lambda_i}\left[\frac{1}{\lambda}+\frac{1}{\mu}\right]^2 - \frac{1}{\lambda\mu}\right). \tag{3.23}$$

Finally, the expectation $\mathbb{E}[Z_jT_j]$ is computed to be

$$\mathbb{E}[Z_jT_j] = \frac{1}{\lambda\mu}\left(\frac{\lambda-\lambda_i}{\lambda_i}\right) + \frac{1}{\lambda(\lambda+\mu)} + \frac{\lambda+2\mu}{(\lambda+\mu)^2\mu}. \tag{3.24}$$

Using these results and substituting in equation (3.21) we obtain that the average AoI of source i in a shared system is given by

$$\Delta_{i,\text{M/M/1,LCFS,preempt}} = \frac{\lambda}{\lambda_i}\left(\frac{1}{\lambda}+\frac{1}{\mu}\right). \tag{3.25}$$

Note that if we fix $(\lambda - \lambda_i)$ and let $\lambda_i \to \infty$, the average age Δ_i tends to $1/\mu$. Thus, as the arrival rate of source i increases and the rates of sources other than i are kept fixed, the average age of source i converges to the average packet service time. Similarly, for N sources and $\lambda_1, \lambda_2, \ldots, \lambda_N \to \infty$, with $\lambda_u = \lambda_v \; \forall \; u$ and v, the average age Δ_i of each source tends to N/μ. Numerical results of the LCFS discipline with preemption will be presented in comparison with other system models in Subsection 3.5.

Gamma distributed service times

Up to this point we have modeled status update systems assuming exponentially distributed interarrival and service times. An interesting alternative studied in [40] is to model the service times as gamma distributed random variables. The gamma distribution is a reasonable

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

approximation for models capturing relay networks [44]. In a relay network the source and the destination are interconnected by means of a number of relays. In the studied case, the transmission time for each hop is exponentially distributed, thus the total transmission time is the sum of i.i.d. exponentials which leads to a gamma distribution.

A sequence of random variables $S_n \sim \Gamma(k_n, \theta_n)$ converges in distribution to a deterministic variable D as k becomes very large. That is, for $\mathbb{E}(S_n) = \frac{1}{\mu}$ and some $\mu > 0$,

$$S_n \to D, \quad \text{as} \quad k \to \infty, \tag{3.26}$$

where $D = \frac{1}{\mu}$ with probability 1. This enables the extension of the gamma distributed service times to deterministic service times by letting $k \to \infty$. The analysis of the average AoI assuming gamma distributed service times under the LCFS queue with and without preemption can be found in [40].

One of the key observations of this work is the following. Under LCFS with preemption as k increases the average age increases for all values of λ. On the other hand, under LCFS without preemption as k increases the average age decreases for all values of λ expect the ones close to zero. Overall, for both deterministic and gamma distributed service times the best strategy is not to preempt especially in a system with high arrival rate.

3.3 Peak Age of Information

In addition to AoI, a new metric called peak age is proposed to serve two objectives. First, depending on the application, it may be a suitable alternative to age and average age. Second, it is a more tractable solution in the analysis of complicated models.

Consider a system consisting of a source-destination link, as shown in Figure 2.1. Observing the peak values in the sawtooth curve we characterize the maximum value of the AoI immediately before an update is received called the *Peak Age of Information* (PAoI) [11], [12]. The *peak age* can be utilized in applications where there is interest in the worst case age or we need to apply a threshold restriction on age. This metric can be used instead of AoI, with the advantage of a simpler formulation, as it will be presented.

Definition 3.3.1 (Peak Age of Information)**.** Let Y_i be the interarrival time of the ith update, and T_i be the corresponding system time. Then, the *Peak Age of Information* (PAoI) metric is defined as the value of *age* achieved immediately before receiving the ith update

$$A_i = Y_i + T_i. \tag{3.27}$$

Figure 2.2 shows an illustrative example of the evolution of AoI as a function of time t. The values of peak age are depicted with A_i. Note that peak age is a discrete stochastic process that takes values at the time instances that belong to the set $\mathcal{I} \doteq \{t'_1, t'_2, \dots t'_{N(\mathcal{T})}\}$.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

In analogy to the AoI, we are interested in small average PAoI in order to maintain fresh information.

Definition 3.3.2 (Time average Peak Age of Information)**.** Suppose that our interval of observation is $(0, \mathcal{T})$. The time average peak age of a status update system is

$$A_{\mathcal{T}} = \lim_{\mathcal{T}\to\infty} \frac{1}{\mathcal{T}} \sum_{i=1}^{N(\mathcal{T})} A_i, \tag{3.28}$$

where $N(\mathcal{T})$ is the number of samples that completed service by time $\mathcal{T}$.

Using the definition (3.27), the *average Peak Age of Information* in the status update system of Figure 2.1 is given by

$$A = \mathbb{E}[Y + T] = \mathbb{E}[Y + W + S], \tag{3.29}$$

where $\lambda = 1/\mathbb{E}[Y]$ and $\mathbb{E}[\cdot]$ is the expectation operator. We assume ergodicity of the stochastic process $\Delta(t)$ but we make no assumptions regarding the distribution of the random variables Y and T, or the service policy. This result also holds when the system is shared among multiple traffic streams.

Comparing the average AoI in (2.12) and the average PAoI of (3.29), we have

$$\begin{aligned} \Delta - A &= \lambda \left(\mathbb{E}[YT] + \mathbb{E}[Y^2]/2 - \mathbb{E}[Y]\mathbb{E}[Y + T] \right) \\ &= \lambda \left(\mathbb{E}[YT] - \mathbb{E}[Y]\mathbb{E}[T] + \mathbb{E}[Y^2]/2 - \mathbb{E}[Y]^2 \right). \end{aligned} \tag{3.30}$$

The difference in (3.30) is a way of estimating how close the two metrics are to each other. The authors in [22] derive exact PAoI expressions for the M/G/1 and the M/G/1/1 models and they show that PAoI serves as an upper bound for AoI. The following lemma shows that PAoI approximates AoI for general G/G/1 models.

Lemma 3.3.1. For a G/G/1 model, we have that

$$A - \frac{3\lambda\mathbb{E}[Y^2]}{2} - \lambda\mathbb{E}[Y]^2 \leq \Delta \leq A + \lambda\mathbb{E}[Y^2]/2. \tag{3.31}$$

Proof. See details in [22]. □

Next, for completeness we provide the average peak age for the M/G/1 as well as the M/M/1 system model. The average PAoI for an M/G/1 system is given by

$$A_{\mathrm{M/G/1}} = \mathbb{E}[Y] + \mathbb{E}[S] + \frac{\lambda\mathbb{E}[S^2]}{2(1-\rho)}. \tag{3.32}$$

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

The result derives from (3.29) and the P-K formula for the system time T provided in [6]. Finally, the average PAoI for an M/G/1/1 system is given by

$$A_{\mathrm{M/G/1/1}} = \mathbb{E}[S] + \mathbb{E}[Y] + \mathbb{E}[Y]\rho. \tag{3.33}$$

The peak age is a suitable metric in applications that impose a threshold restriction on age.

In the following subsection we introduce the reader to the notion of obsoleteness and present the first analysis of a system with multiple servers. The need for treating waste of resources becomes apparent and leads to packet management techniques in Subsection 3.5, where PAoI is used to provide tractable analysis.

3.4 Availability of Resources

After the introductory models in Section 2, some works departed from the simple model of the M/M/1 to more elaborated ones that capture the availability of more resources in a network. In this setup, we take into consideration a more dynamic feature of networks, that is, packets traveling over a network might reach the destination through different paths thus the delay of each packet might differ. In this case, we say that the source and the destination are separated by a *network cloud* as depicted in Figure 2.1.

We assume that status updates are instantaneously served upon generation, which means that there is no buffer between the source and the network cloud, nor delays due to waiting times. Following our previous definitions, where the system time of update i is $T_i = W_i + S_i$, we now refer to service time of update i as $S_i = t'_i - t_i$, for $W_i = 0\,\forall i$. The service times are modeled as i.i.d. exponentially distributed random variables with mean $1/\mu$. The random network delay is a simplified model that captures the possibility of packets arriving at the destination out of order, due to various dynamics of the network, such as link scheduling, competing data traffic, or multiple paths.

Definition 3.4.1 (Informative packet)**.** Consider a system consisting of a source-destination communication link, separated by a network cloud. Assume Y_i and S_i are the interarrival and service times of update i, respectively. Then, we define an *informative* packet as a packet that carries the newest information compared to the packets arriving at the destination prior to it. Mathematically, the condition for a packet m being informative is

$$S_m < S_r + \sum_{a=m+1}^{r} Y_a,\ \forall\, r > m. \tag{3.34}$$

Definition 3.4.2 (Obsolete packet)**.** A packet l is said to be *obsolete* if there is at least one (packet with) $k \geq 1$ generated after l, such that $t'_l > t'_{l+k}$. An informative packet is one that is not rendered obsolete.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

The model described can be viewed as a system with infinite memoryless servers with the Kendall notation M/M/∞. However, modeling the network as M/M/∞ does not reflect the fact that packets entering the system first, most likely will finish service first. This is the case with a single server queue, where we consider in-order reception. On the other hand, a single server system does not reflect the dynamics of a network (e.g., changing topology) which might cause out-of-order reception of packets. Therefore, we are interested in investigating different systems with two or more servers which reflect both the transmission path diversity and the service priority principles [27]. A suitable combination of queueing and out-of-order reception can model various network topologies. Towards this direction, this subsection compares the performance of the M/M/1, M/M/2, and M/M/∞ cases and demonstrates the tradeoff between the AoI at the destination and the waste of network resources as the number of servers varies.

3.4.1 The M/M/∞ system model

We consider the system model described in this subsection consisting of a source that transmits packets through a network with infinity numbers of servers. For modeling purposes a server can be viewed as a wireless channel in a queue theoretic setup. Status updates are generated according to a Poisson process with mean λ and the service times of each server are modeled as exponential random variables with mean $1/\mu$. We begin with a preliminary analysis of the probability of possible events within this framework. Next, we present a closed form expression for the average AoI of this system.

We start by denoting $\mathcal{E}(n)$ the event that a packet is informative and that it renders n other packets obsolete. Let $E_1(i)$ be the event that a packet is informative and $E_2(n)$ be the event that a packet renders exactly n of the previous packets obsolete, i.e., arrives before the previous n packets and after the $(n+1)$st previous packet. Eventually, we have $\mathcal{E}(n) = E_1(i) \cap E_2(n)$. Note, that the steady-state probability of $\mathcal{E}(n)$ is independent of i.

First, the conditional probability of $E_1(i)$ given its service time S_i and the interarrival times Y_{i+1}^{∞} of future packets, are derived. We continue by deriving the conditional probability of $E_2(n)$ given the service time S_i and the interarrival times of the previous n packets Y_{i-n}^{i}, and then averaging over all the y_{i-n}^{i} to obtain $\Pr(E_2(n)|S_i = s_i)$. Observe that the probability of the intersection of the events $E_1(i)$ and $E_2(n)$ is equal to the probability of their product since they are conditionally independent given S_i. This is true because $E_1(i)$ depends on S_i^{∞} and Y_{i+1}^{∞}, while $E_2(n)$ depends on S_{i-n-1}^{i} and Y_{i-n}^{i}, and all the interarrival and service times are independent. Finally, we obtain (see details in [27])

$$\Pr(\mathcal{E}(n)) = \Pr(E_1(i) \cap E_2(n)) = \frac{\lambda^n \mu}{\prod_{k=1}^{n+1}(\lambda + k\mu)}. \tag{3.35}$$

To compute the average status age, we also need the statistics of the interarrival and service times conditioned on the event $\mathcal{E}(n)$. For completeness we provide all the terms the average

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

status age consists of. We start with the calculation of the conditional expectation of the service time of update i, $\mathbb{E}[S_i|\mathcal{E}(n)]$, that is given by

$$\mathbb{E}[S_i|\mathcal{E}(n)] = \frac{1}{\lambda+(n+1)\mu}\left(1+\frac{\lambda}{\lambda+(n+2)\mu}\right). \tag{3.36}$$

Following the same methodology we derive useful conditional expectations related to Y_{i-n}^{i} [27]. The conditional sum of means is

$$\mathbb{E}\left[\sum_{k=i-n}^{i} Y_k|\mathcal{E}(n)\right] = \sum_{k=0}^{n}\frac{1}{\lambda+k\mu}+\frac{n+1}{\lambda}\sigma(n), \tag{3.37}$$

where

$$\sigma(n) = \sum_{r=1}^{\infty}\left[\frac{\lambda^r}{(n+r+1)\prod_{k=1}^{r}(\lambda+(n+k)\mu)}\left(1-\frac{\lambda}{\lambda+(n+r+1)\mu}\right)\right].$$

The conditional sum of second moments can be similarly derived

$$\mathbb{E}\left[\sum_{k=i-n}^{i} Y_k^2|\mathcal{E}(n)\right] = \sum_{k=0}^{n}\frac{2}{(\lambda+k\mu)^2}+\frac{2(n+1)}{\lambda^2}\Big(1+\frac{\lambda}{\lambda+(n+1)\mu}\Big)\sigma(n). \tag{3.38}$$

Finally, the conditional sum of crossterms is given by

$$\begin{aligned}&\mathbb{E}\left[\sum_{j=i-n}^{i-1}\sum_{k=j+1}^{i} 2Y_jY_k|\mathcal{E}(n)\right] = \\ &= \sum_{j=0}^{n-1}\sum_{k=j+1}^{n}\frac{2}{(\lambda+j\mu)(\lambda+k\mu)}+\frac{(n+1)\sigma(n)}{\lambda}\sum_{k=1}^{n}\frac{2}{\lambda+k\mu}.\end{aligned} \tag{3.39}$$

Before the calculation of the average status age for the M/M/∞ model the derivation of the probability of a packet rendered obsolete which is equal to the probability of a packet not being informative is presented. The result can be used as an indicator of the performance of the system, since obsolete packets correspond to waste of resources. Thus, it is meaningful to minimize the percentage of obsolete packets among the transmitted packets. The probability of a packet i becoming obsolete is

$$\begin{aligned}&1-\Pr\{E_1(i)\} = \\ &= \frac{\rho}{\rho+1}-\sum_{r=1}^{\infty}\frac{\rho^r}{(r+1)\prod_{k=1}^{r}(\rho+k)}\left(1-\frac{\rho}{\rho+r+1}\right),\end{aligned} \tag{3.40}$$

where ρ is the server utilization. This expression indicates that the probability of a packet rendered obsolete is solely a function of ρ.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

Average age analysis derivation and bounds

In the M/M/∞ case, the key element compared to the previously discussed models is that some of the interarrival times Y_i might correspond to non-informative packets, therefore for each informative packet we should also consider the previous n packets rendered obsolete. As a result we have

$$\frac{1}{\mathbb{E}[Y]} = \lambda \Pr(\mathcal{E}(n)),$$

$$\mathbb{E}[Y^2]/2 = \left(\mathbb{E}\left[\sum_{k=i-n}^{i} Y_k^2 | \mathcal{E}(n)\right] + \mathbb{E}\left[\sum_{j=i-n}^{i-1}\sum_{k=j+1}^{i} 2Y_jY_k | \mathcal{E}(n)\right]\right) \Big/ 2,$$

$$\mathbb{E}[YS] = \mathbb{E}\left[\sum_{k=i-n}^{i} Y_k | \mathcal{E}(n)\right] \mathbb{E}[S_i | \mathcal{E}(n)].$$

Substituting equations (3.35), (3.36), (3.37), (3.38), and (3.39) to (2.12) we obtain

$$\Delta_{\mathrm{M/M/\infty}} = \lambda \sum_{n=0}^{\infty} \Pr(\mathcal{E}(n))$$
$$\left[\sum_{j=0}^{n}\left(\frac{1}{\lambda+j\mu}\left(\frac{2\lambda+(n+2)\mu}{(\lambda+(n+1)\mu)(\lambda+(n+2)\mu)} + \sum_{k=j}^{n}\frac{1}{\lambda+k\mu}\right)\right) + \right.$$
$$\left.\frac{(n+1)\sigma(n)}{\lambda}\left(\frac{2\lambda+(n+2)\mu}{(\lambda+(n+1)\mu)(\lambda+(n+2)\mu)} + \sum_{l=0}^{n+1}\frac{1}{\lambda+l\mu}\right)\right]. \tag{3.41}$$

It is apparent that the exact analysis of AoI becomes harder as the model becomes more complicated. Due to this complexity we consider upper and lower bounds. A simple upper bound is to compute the average of the trapezoid areas for all packets, whether or not they are informative. Therefore, the bottom edge of the trapezoids consists of only one interarrival time (either of an informative or an obsolete packet) and we obtain

$$\Delta_{UB,\mathrm{M/M/\infty}} = \lambda\left(\mathbb{E}[Y]\mathbb{E}[S] + \mathbb{E}[Y^2]/2\right) = \frac{1}{\lambda} + \frac{1}{\mu}. \tag{3.42}$$

A lower bound can be computed as follows; consider that the service time of a packet can not be greater than the interarrival time of the next generated packet. Conditioning on the probability that $S_\alpha < Y_{\alpha+1}$ we assume that a new status update is generated after the previous update finishes service. This assumption is similar to the just-in-time bound on Subsection 2.2 where the server state is considered known. The area under the sawtooth of $\Delta(t)$ is now smaller than that of the original system and thus the lower bound is obtained

$$\Delta_{LB,\mathrm{M/M/\infty}} = \frac{1}{\lambda} + \frac{1}{\lambda+\mu} - \frac{\lambda\mu}{(\lambda+\mu)^3}. \tag{3.43}$$

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

The bounds are tighter for small values of the server utilization ρ and they become looser as ρ increases. This is more evident for smaller values of μ where the effect of the assumptions is stronger [27].

In [27] the expression (3.41) of the average age for the M/M/∞ model has been evaluated as shown in Figure 3.5. The theoretical and simulated status age is plotted versus the utilization ρ for $\mu = 0.5, 1, 1.5$. The upper and lower bounds $\Delta_{UB,\mathrm{M/M/\infty}}$ and $\Delta_{LB,\mathrm{M/M/\infty}}$ are also depicted with dotted and dashed lines, respectively. We observe that the average age is monotonically decreasing as the server utilization increases, since more frequent transmissions lead to a more updated destination node. However, this comes with the cost of wasted network resources due to obsolete packets as we will see in detail in Section 3.4.3.

3.4.2 The M/M/2 system model

In this subsection we study a status update system consisting of two servers capturing both the possibility of packets arriving at the destination out of order and the effect of queuing. More specifically, we assume that the interarrival times Y_i and the service times S_i are exponentially distributed with mean $1/\lambda$ and $1/\mu$, respectively. We will refer to this model as M/M/2 and we will present the average status age analysis.

In this setup, although we have in order queueing we might have out-of-order reception. Packets are served with a first-come-first-served discipline and if upon generation they find both servers busy they have to wait in the queue, and the waiting time is W_i. Furthermore, assume that the system is empty and a packet arrives and enters into the first server. Then, a second packet arrives and finds the first server busy, so it enters into the second server. If packet 2 finished service before packet 1, it means that it renders packet 1 obsolete since packet 2 is the most recently generated packet.

Note that for a two-server system, two consecutive packets cannot be made obsolete, since this would mean that a packet that enters the system after them will complete service first. This, however, is not possible for an M/M/2 system with a FCFS, because two packets will occupy both servers, and one of them must complete service prior to any future packets entering service. We denote $\tilde{Y}_i$ the interarrival time and $\tilde{T}_i$ the system time of the ith informative packet. Observe that $\tilde{Y}_i$ can consist of either one or two interarrival times of generated packets, while $\tilde{T}_i$ is equal to $\tilde{W}_i + \tilde{S}_i$. After applying expression (2.7), the average age can be therefore written as

$$\Delta_{\mathrm{M/M/2}} = \frac{\mathbb{E}[\tilde{Y}\tilde{S}] + \mathbb{E}[\tilde{Y}^2]/2}{\mathbb{E}[\tilde{Y}]} = \frac{\mathbb{E}[\tilde{W}\tilde{Y}] + \mathbb{E}[\tilde{Y}]\mathbb{E}[\tilde{S}] + \mathbb{E}[\tilde{Y}^2]/2}{\mathbb{E}[\tilde{Y}]}. \tag{3.44}$$

We now proceed by categorizing the informative packets of an M/M/2 system into two different types. The event A that defines an informative packet of type α, is the

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

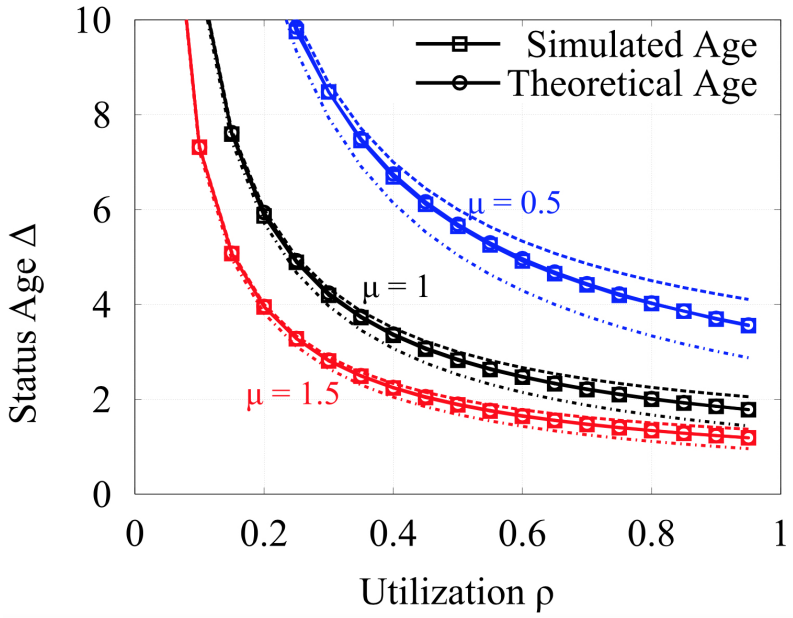

Figure 3.5: Average status age vs server utilization for the M/M/∞ queue [27].

case where an informative packet is preceded by another informative packet. In the case where informative packets are separated by an obsolete packet, event B, we categorize the informative packet that renders the previous packet obsolete as of type b. For the first type, the interarrival time $\tilde{Y}_\alpha$ consists solely of a single interarrival time, whereas for type b, $\tilde{Y}_b$ consists of exactly two typical interarrival times, one of an informative and one of an obsolete packet (i.e. $\tilde{Y}_i = Y_i + Y_{i+1}$). Encountering both packets of type α and b, we cover the whole sample space of packets that contribute to the Age of Information of the system at the destination. Thus, we have

$$\Delta_{\mathrm{M/M/2}} = \lambda \Big[p_\alpha \big(\mathbb{E}[\tilde{W}_\alpha \tilde{Y}_\alpha] + \mathbb{E}[\tilde{Y}_\alpha]\mathbb{E}[\tilde{S}_\alpha] + \mathbb{E}[\tilde{Y}_\alpha^2]/2\big) + p_b \big(\mathbb{E}[\tilde{W}_b \tilde{Y}_b] + \mathbb{E}[\tilde{Y}_b]\mathbb{E}[\tilde{S}_b] + \mathbb{E}[\tilde{Y}_b^2]/2\big) \Big], \tag{3.45}$$

where p_α is the probability of event A and p_b is the probability of event B.

The expected value $\mathbb{E}[\tilde{W}_\alpha \tilde{Y}_\alpha]$ can be derived using iterated expectation, i.e.,

$$\mathbb{E}[\tilde{W}_\alpha \tilde{Y}_\alpha] = \int_0^\infty y\, \mathbb{E}[W_i | Y_i = y] f_{Y_{i/\alpha}}(y)\, \mathrm{d}y. \tag{3.46}$$

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

Similarly, for informative packets of type b we have

$$\mathbb{E}[\tilde{W}_b\tilde{Y}_b] = \int_0^\infty \int_0^\infty (y_1+y_2)\,\mathbb{E}[W_i|Y_{i-1}=y_1, Y_i=y_2] \times f_{Y_{i-1/b},Y_{i/b}}(y_1,y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2. \tag{3.47}$$

The derivation details of equations (3.46) and (3.47) can be found in [26] and [27]. Note that the distribution of the interarrival times and the system times for a packet of type α or b are not the same with those of a typical packet. Computation of $f_{Y_{i/\alpha}}$ and $f_{Y_{i-1/b},Y_{i/b}}$ requires knowledge of the joint distribution of Y and T, which is difficult to derive. A way to proceed is conditioning on events that can be computed in a straightforward manner using the memoryless property of the exponential interarrival and service times. The probabilities of events of type A are given in [27, pp. 1365-1366] and the probabilities of events of type B are given in [27, pp. 1366-1367]. The probabilities p_α and p_b can be then derived by taking the sum of the probability of events of type A and B, respectively. The probability that a packet is rendered obsolete is the probability that a packet is not informative, which is equal to $1-(p_\alpha+p_b)$.

Average age analysis: approximations and bounds

The analysis of the average AoI for the M/M/2 queue is clearly difficult, thus here we will proceed by providing approximations and bounds. We let the distribution of the interarrival time of an informative packet of type α to be an exponential interarrival time Y_i, while for a packet of type b, the approximated interarrival time is the sum of two i.i.d. exponential interarrival times. As another approximation, we consider that $\mathbb{E}[\tilde{S}]$ is the same for both type α and b packets. Then, the approximate average age can be computed as (see details in [27])

$$\Delta_{\mathrm{M/M/2}} \approx \lambda\Big[p_\alpha\big(\mathbb{E}[\tilde{W}\tilde{Y}_\alpha] + \mathbb{E}[\tilde{Y}_\alpha]\mathbb{E}[\tilde{S}] + \mathbb{E}[\tilde{Y}_\alpha^2]/2\big) + p_b\big(\mathbb{E}[\tilde{W}\tilde{Y}_b] + \mathbb{E}[\tilde{Y}_b]\mathbb{E}[\tilde{S}] + \mathbb{E}[\tilde{Y}_b^2]/2\big)\Big]. \tag{3.48}$$

In addition, a simple upper bound is obtained by encountering all transmitted packets as informative, similar to the M/M/∞ case. This upper bound is given by

$$\Delta_{\mathrm{UB,\ M/M/2}} = \frac{1}{\mu}\left(1+\frac{1}{2\rho}+\frac{2\mu\rho^3}{(1+\rho)(1-\rho)}\right). \tag{3.49}$$

For the lower bound, assume that the interarrival time is the same with a single typical interarrival time, and argue that the interarrival time of an informative packet is stochastically greater than or equal to that of a typical packet. Let $E_3(i)$ be the event that packet i

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

is informative in an M/M/2 system. In [27] it is shown that $\Pr(Y_i > y|E_3(i)) \geq \Pr(Y_i > y)$ holds, thus, a lower bound can be given by

$$\Delta_{\text{LB, M/M/2}} = \tilde{\lambda}(\mathbb{E}[\tilde{W}\tilde{Y}_\alpha] + \mathbb{E}[\tilde{Y}_\alpha]\mathbb{E}[\tilde{S}] + \mathbb{E}[\tilde{Y}_\alpha^2]/2), \tag{3.50}$$

where $\tilde{Y}_\alpha$ is equal to a typical interarrival time and $\tilde{\lambda} = \lambda(p_\alpha + p_b)$.

In Figure 3.6, we evaluate the performance of the M/M/2 model by comparing the average simulated age with the approximation and the upper and lower bounds as a function of the server utilization ρ, for $\mu = 1$ [27]. Additionally, the age for the M/M/1 model is plotted as a point of reference. We observe that the approximated age is very close to the simulated age. Furthermore, the upper and lower bounds are relatively tight for smaller ρ, but they start deviating as ρ increases, especially in the upper bound case. As ρ approaches 1, the number of obsolete packets increases, thus, we encounter more of them as informative and the gap between the theoretical age and its upper bound increases. For the lower bound, we assume that an informative interarrival time is stochastically the same as a typical interarrival time, therefore as ρ increases informative interarrival times are more likely to consist of two typical interarrival times. Finally, we can conclude that the status age for the M/M/1 model is almost twice of that for the M/M/2 model.

3.4.3 Tradeoff between status age and obsolete packets

Figure 3.7 depicts the average status age for the M/M/1, M/M/2 and M/M/∞ models versus the arrival rate λ, when the service rate is $\mu = 1$. Note that as ρ increases, the number of packets in the system P increases as well, and as $\rho \to 1$, we have $P \to \infty$. If $\rho > 1$, the server cannot sustain the arrival rate and the queue length increases without bound. Therefore, for M/M/c, c=1, 2, the age also approaches infinity as λ approaches c. Overall, increasing the number of servers in the system, results in decreased status age.

Figure 3.7 also depicts the percentage of packets that are rendered obsolete. For the M/M/1 model, it is not possible for packets to arrive out of order, thus there are no obsolete packets. From the M/M/2 and M/M/∞ cases we see that more servers lead to more obsolete packets, since more packets are served simultaneously. The utilization of more servers reduces the average status age, but this comes at the cost of wasted network resources in heavy loads.

3.5 Packet Management

To avoid congestion in networks, packet management techniques or flow control can be utilized in order to manage the traffic entering a network. The focus of this subsection is on packet management by dropping or replacing packets. In this way the network can

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

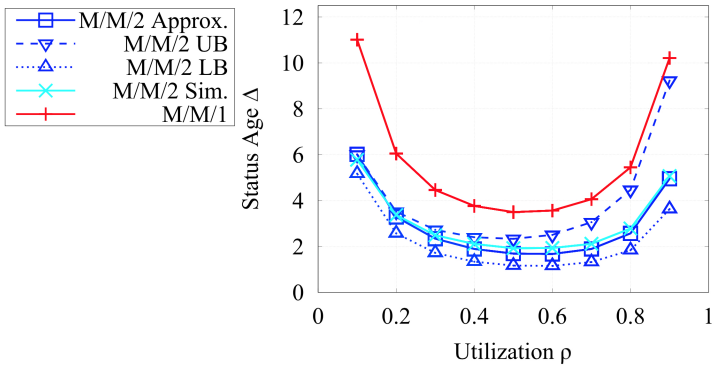

Figure 3.6: Average status age vs server utilization of the M/M/2 and M/M/1 queues, for $\mu = 1$ [27].

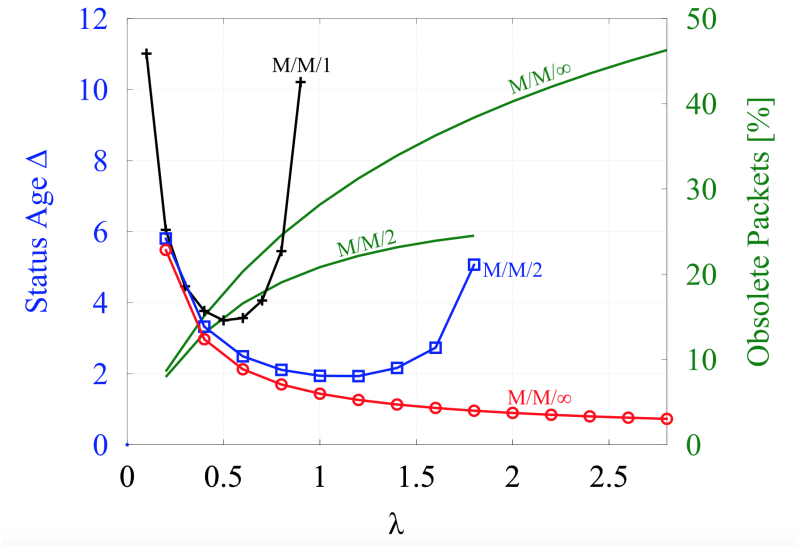

Figure 3.7: Average status age and % of obsolete packets vs arrival rate for the M/M/1, M/M/2, and M/M/∞ queues, for $\mu = 1$ [27].

utilize the resources more efficiently and reduces the imposed packet delay thus, the Age of Information performance can be improved.

In this subsection, we consider a system with a single or multiple independent sources that can discard some of the generated packets. The selection process of packets to discard is referred to as *packet management* [11], [12]. The packet management can improve the performance of the system with respect to the staleness of the transmitted information, similar to the case with the LCFS queue discipline. We consider three packet management policies and we compare the results of the average age. For all policies we assume Poisson

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

arrivals, an FCFS discipline and exponentially distributed service times.

1) For the first policy, we assume that a packet which arrives while another packet is being served is discarded, and packets that find the server idle immediately receive service. This means that no packets are kept in the queue waiting for transmission. The described policy is modeled as an M/M/1/1 queue, where the last entry in the Kendall notation refers to the total capacity of the queue, which is a single packet in service.

2) The second policy, assumes that a single packet may be kept in the queue waiting for transmission if another packet is being served at the same time. A packet which finds both the server and the single queue position occupied is discarded. A packet that finds the server idle is immediately receiving service. Here the capacity of the system model is one packet in queue and one packet in service. Thus, the described policy is modeled as an M/M/1/2 queue.

3) The third packet management policy also assumes that a packet which arrives while another packet is being served may be kept in the queue waiting for transmission. However, the packets waiting for transmission are replaced by newly generated packets. This model will be identified as M/M/1/2* [12]. Note that this is a non-traditional queueing model, for which some of the classic results from queueing theory, such as Little's law, do not apply [35]. Furthermore, we discuss the extension of this policy for a system with N independent sources provided in [45]. In this case, we maintain a queue with only the latest status packet of each source by replacing any previously queued packet from that source. We refer to this policy as M/M/1/(N+1)*.

Next, an illustrative example is presented in order to highlight the effect of packet management policies on the packets that receive service and the effect on the age. Recall that the system time of update i is denoted by T_i and the interdeparture time of update i is Z_i. In Figure 3.8, for the M/M/1/1 queue, the packets that arrive at times $t_\star$ and $t_{\star\star}$ are discarded, since they find the server busy. For the M/M/1/2 queue, age is depicted in Figure 3.9 where the packet generated at time $t_\star$ finds the server busy and the queue occupied thus it is discarded. Figure 3.10 shows how M/M/1/2* responds in this scenario of arrivals and we see that the packet generated at time $t_\star$ spends some time in the queue waiting for transmission and then is replaced by a new packet arriving at time t_4. Since no packets arrive between the interval t_4 and t_3', the new packet is served. Note that the subindex i is used to refer to successfully transmitted packets only. Packets generated between t_i and t_{i+1} include packets that do not receive service.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

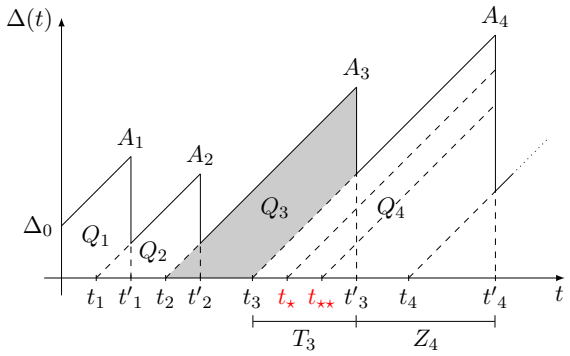

Figure 3.8: Example of age evolution for an M/M/1/1 system [12].

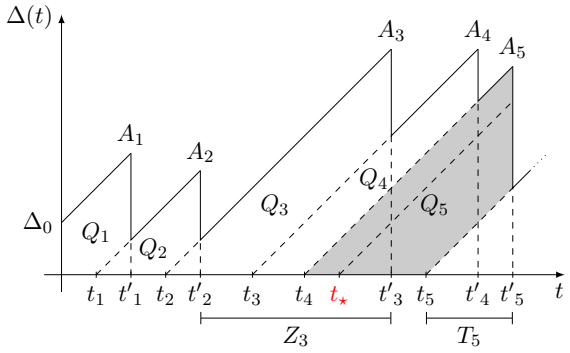

Figure 3.9: Example of age evolution for an M/M/1/2 system [12].

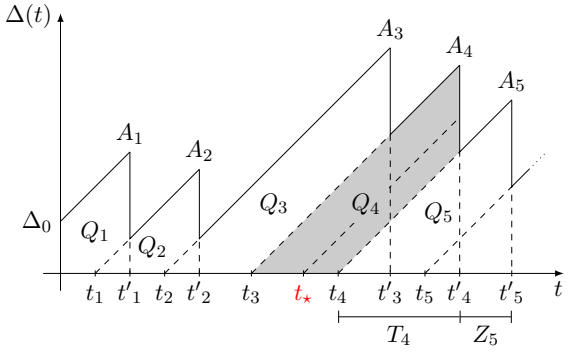

Figure 3.10: Example of age evolution for an M/M/1/2* system [12].

Preliminary analysis

To derive the average age and peak age for the three packet management policies we revisit equations (2.12) and (3.29) as in [12]. In Figures 3.8, 3.9, and 3.10 we observe that the

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

trapezoid areas Q_i can be calculated as

$$\begin{aligned}\mathbb{E}[Q_i] &= \frac{1}{2}\mathbb{E}[(T_{i-1}+Z_i)^2] - \frac{1}{2}\mathbb{E}[T_i^2] \\ &= \mathbb{E}[Z_i T_{i-1}] + \mathbb{E}[Z_i^2]/2, \quad k \geq 3,\end{aligned} \tag{3.51}$$

where the second equality follows from the fact that T_{i-1} and T_i are identically distributed for $k \geq 3$. This alternative definition allows us to rewrite the average age as

$$\Delta = \lambda_e \mathbb{E}[Q] = \lambda_e(\mathbb{E}[ZT] + \mathbb{E}[Z^2]/2), \tag{3.52}$$

where Y, T, Z denote the interarrival, system, and interdeparture times. The effective arrival rate is λ_e, and $\mathbb{E}[\cdot]$ is the expectation operator. Furthermore, the peak Age of Information is redefined as

$$A_i = Z_i + T_{i-1}, \tag{3.53}$$

thus, yields the *average Peak Age of Information*

$$A = \mathbb{E}[Z+T] = \mathbb{E}[Z+W+S], \tag{3.54}$$

Remark 3.5.1. The results from equation (3.52) and equation (3.54) holds when the system is shared among multiple traffic streams.

We note that the random variables Z and T are dependent and this complicates the calculations of the average age in the general case, since we are not aware of their joint distribution. Nonetheless it is possible to describe an event such that Z_i and T_{i-1} are conditionally independent.

We denote ψ_i the event that the system is empty after the ith successful transmission. Under the assumption of Poisson arrivals, the remaining time until the next arrival is exponentially distributed with mean $1/\lambda$. Then, the arriving packet will be served immediately, and completes service after an exponentially distributed period with mean $1/\mu$. Further, let $\bar{\psi}_i$ be the complementary event that the ith packet left behind a system with a packet waiting in the queue. In this case, the first packet in the queue will receive service immediately after the departure of packet i and it will complete service after an exponentially distributed service time.

The conditional distribution of the interdeparture time Z_i is independent of the queuing model we consider. Given the event ψ_{i-1}, the distribution of Z_i is given by the convolution of the distribution of Y_i and S_i, which yields

$$f(z|\psi_{i-1}) = \frac{\lambda\mu}{\mu-\lambda}\left[e^{-\lambda z} - e^{-\mu z}\right], \tag{3.55}$$

$$\mathbb{E}[Z_i|\psi_{i-1}] = \frac{1}{\lambda} + \frac{1}{\mu}, \tag{3.56}$$

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

$$\mathbb{E}[Z_i^2|\psi_{i-1}] = \frac{2(\lambda^2+\lambda\mu+\mu^2)}{\lambda^2\mu^2}. \tag{3.57}$$

The interdeparture time Z_i, conditioned on $\bar{\psi}_i$, coincides with the service time S_i, that is

$$f(z|\bar{\psi}_{i-1}) = \mu e^{-\mu z}, \tag{3.58}$$

$$\mathbb{E}[Z_i|\bar{\psi}_{i-1}] = \frac{1}{\mu}, \tag{3.59}$$

$$\mathbb{E}[Z_i^2|\bar{\psi}_{i-1}] = \frac{2}{\mu^2}. \tag{3.60}$$

The variables Z_i and T_{i-1} are conditionally independent under the event ψ_{i-1}. Then, the expectation $\mathbb{E}[Z_iT_{i-1}]$ is calculated as follows

$$\begin{aligned}\mathbb{E}[T_{i-1}Z_i] &= \mathbb{P}(\psi_{i-1})\mathbb{E}[Z_iT_{i-1}|\psi_{i-1}] + \mathbb{P}(\bar{\psi}_{i-1})\mathbb{E}[Z_iT_{i-1}|\bar{\psi}_{i-1}] \\ &= \mathbb{P}(\psi_{i-1})(\mathbb{E}[Z_i|\psi_{i-1}]\mathbb{E}[T_{i-1}|\psi_{i-1}]) \\ &\quad + \mathbb{P}(\bar{\psi}_{i-1})(\mathbb{E}[Z_i|\bar{\psi}_{i-1}]\mathbb{E}[T_{i-1}|\bar{\psi}_{i-1}]).\end{aligned} \tag{3.61}$$

The M/M/1/1, M/M/1/2, and M/M/1/2* models are studied in [12] for a single source system. In this case, the probability that a departure leaves the system empty, is equal to the steady state probability that the system is empty [10, Ch. 5] and it can be calculated for each model separately. Moreover, for each model we will derive the pdf of the peak age by conditioning the pdfs of T_{i-1} and Z_i on the event ψ_{i-1}. The conditional pdf of the peak age then is obtained as the convolution

$$f(\alpha|\psi_{i-1}) = f(t|\psi_{i-1}) * f(z|\psi_{i-1}). \tag{3.62}$$

Following the same methodology for the event $\bar{\psi}_{i-1}$ we get the distribution of the PAoI

$$f(\alpha) = \mathbb{P}(\psi_{i-1})f(\alpha|\psi_{i-1}) + \mathbb{P}(\bar{\psi}_{i-1})f(\alpha|\bar{\psi}_{i-1}). \tag{3.63}$$

In the remaining of the section, the average age and peak age are calculated for each one of the described models. In each case, we need to derive the steady state distribution of the number of packets in the system, in order to calculate the probability $\mathbb{P}(\psi_i)$. In addition, the conditional distribution of the system time T_{i-1} given ψ_{i-1} is needed.

The M/M/1/1 system model

The M/M/1/1 queue can be described through a two-state Markov chain, where each state represents the server being idle or busy. It can be then easily shown that the steady state probabilities of the system are [35, Ch. 3]

$$p_0 = \frac{\mu}{\lambda+\mu} \quad \text{and} \quad p_1 = \frac{\lambda}{\lambda+\mu}. \tag{3.64}$$

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

The probability that the ith packet leaves an empty system upon departure is $\mathbb{P}(\psi_i) = 1$. The system time is equal to the service time which is exponentially distributed with expected value $\mathbb{E}[T_{i-1}] = 1/\mu$. Packets enter the system whenever the server is empty, thus, the effective arrival rate is $\lambda_e = (1/\mathbb{E}[Y])(1 - p_1)$. Substituting (3.57) and (3.61) to (3.52) the average status age Δ is

$$\Delta_{\mathrm{M/M/1/1}} = \frac{1}{\lambda} + \frac{2}{\mu} - \frac{1}{\lambda + \mu}. \tag{3.65}$$

Note that for infinitely large arrival rate λ the average age yields

$$\lim_{\lambda \to \infty} \Delta_{\mathrm{M/M/1/1}} = \frac{2}{\mu}. \tag{3.66}$$

This is equal to the just-in-time lower bound for a queue with an FCFS discipline found in Section 2.2. Therefore, an M/M/1/1 queue with arrivals independent to the current status of the server can perform equally good to a system with just-in-time service-depended arrivals.

For the peak age, equation (3.63) yields

$$\begin{aligned} f(\alpha)_{\mathrm{M/M/1/1}} &= f(\alpha|\psi_{i-1}) = f(t|\psi_{i-1}) * f(z|\psi_{i-1}) \\ &= \left(\frac{\mu}{\lambda - \mu}\right)^2 \lambda(e^{-\lambda\alpha} - e^{-\mu\alpha} + (\lambda - \mu)\alpha e^{-\mu\alpha}). \end{aligned} \tag{3.67}$$

Then, the probability that the peak age surpasses a threshold is

$$\begin{aligned} \mathbb{P}(A > \alpha)_{\mathrm{M/M/1/1}} = \left(\frac{\mu}{\lambda - \mu}\right)^2 e^{-\lambda\alpha} + \left[1 - \left(\frac{\mu}{\lambda - \mu}\right)^2\right] e^{-\mu\alpha} \\ + \frac{\lambda\mu}{\lambda - \mu}\alpha e^{-\mu\alpha}. \end{aligned} \tag{3.68}$$

Finally, the average peak age can be calculated directly as

$$A_{\mathrm{M/M/1/1}} = \mathbb{E}[Z + T] = \frac{1}{\mu} + \frac{1}{\lambda} + \frac{1}{\mu} = \frac{1}{\lambda} + \frac{2}{\mu}. \tag{3.69}$$

The M/M/1/2 system model

The M/M/1/2 queue can be described as a three-state Markov chain, where each state represents an empty system, a single packet receiving service or a system with two packet, one in service and one in the queue. Denoting by $\rho = \lambda/\mu$ the server utilization, it can be shown that the steady state probabilities of the system are given by [35, Ch. 3]

$$p_j = \frac{\rho^j}{1 + \rho + \rho^2}, \quad j \in \{0, 1, 2\}. \tag{3.70}$$

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

Recall that the probability that a departure leaves the system empty is equal to the steady state probability that the system is empty. We normalize the probabilities thus

$$\mathbb{P}(\psi_i) = \frac{p_0}{p_0 + p_1} = \frac{\mu}{\lambda + \mu}, \tag{3.71}$$

$$\mathbb{P}(\bar{\psi}_i) = \frac{p_1}{p_0 + p_1} = \frac{\lambda}{\lambda + \mu}. \tag{3.72}$$

Packets enter the system whenever the server or queue is empty, thus the effective arrival rate is $\lambda_e = (1/\mathbb{E}[Y])(1 - p_2)$.

Now we can calculate the second term in (3.52), that is

$$\begin{aligned}\frac{1}{2}\mathbb{E}[Z_i^2] &= \frac{1}{2}\left(\mathbb{P}(\psi_{i-1})\mathbb{E}[Z_i^2|\psi_{i-1}] + \mathbb{P}(\bar{\psi}_{i-1})\mathbb{E}[Z_i^2|\bar{\psi}_{i-1}]\right) \\ &= \frac{1}{2}\left(\frac{2(\lambda^2 + \lambda\mu + \mu^2)}{(\lambda\mu)^2}\frac{\mu}{\lambda + \mu} + \frac{2}{\mu^2}\frac{\lambda}{\lambda + \mu}\right) = \frac{1}{\lambda^2} + \frac{1}{\mu^2}.\end{aligned} \tag{3.73}$$

Next, we characterize the conditional distribution of the system time T_{i-1} given the events ψ_{i-1} and $\bar{\psi}_{i-1}$. In accordance to these derivations, the conditional expectations of T_{i-1} are obtained in [12] as

$$\mathbb{E}[T_{i-1}|\psi_{i-1}] = \frac{1}{\mu}, \tag{3.74}$$

$$\mathbb{E}[T_{i-1}|\bar{\psi}_{i-1}] = \frac{2}{\mu}. \tag{3.75}$$

Substituting the probabilities (3.71), (3.72), the conditional expectations of the system time (3.74), (3.75), and the conditional expectations of the interdeparture time (3.56), (3.59), to (3.61) we obtain

$$\mathbb{E}[T_{i-1}Z_i] = \frac{2\lambda^2 + \lambda\mu + \mu^2}{\lambda\mu^2(\lambda + \mu)}. \tag{3.76}$$

Finally, using (3.73) and (3.76) the average AoI is calculated as

$$\Delta_{\mathrm{M/M/1/2}} = \frac{1}{\lambda} + \frac{3}{\mu} - \frac{2(\lambda + \mu)}{\lambda^2 + \lambda\mu + \mu^2}. \tag{3.77}$$

For infinitely large arrival rate λ, the average age yields

$$\lim_{\lambda\to\infty} \Delta_{\mathrm{M/M/1/2}} = \frac{3}{\mu}. \tag{3.78}$$

This value is greater that the just-in-time lower bound for a queue with an FCFS discipline described in Section 2.2 and the M/M/1/1 case. Intuitively, for large values of the arrival rates it is better to discard the packets in the queue, since queueing would cause an increase in the information age.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

Using the same approach as in the M/M/1/1 case, the probability the peak age surpasses a threshold is

$$\mathbb{P}(A > \alpha)_{\mathrm{M/M/1/2}} = \frac{\mu^3}{(\lambda-\mu)^2(\lambda+\mu)} e^{-\lambda\alpha} + \frac{\lambda}{2(\lambda-\mu)^2(\lambda+\mu)} e^{-\mu\alpha} \times \left[\mu^2\alpha^2(\lambda-\mu)^2 + 2\lambda\mu(\lambda-\mu)\alpha + 2(\lambda^2-\lambda\mu-\mu^2)\right]. \tag{3.79}$$

Finally, the average peak age can be calculated directly as

$$A_{\mathrm{M/M/1/2}} = \mathbb{E}[Z+T] = \frac{1}{\lambda} + \frac{\lambda}{\mu(\lambda+\mu)} + \frac{\mu+2\lambda}{\mu(\lambda+\mu)} = \frac{1}{\lambda} + \frac{3}{\mu} - \frac{2}{(\lambda+\mu)}. \tag{3.80}$$

The M/M/1/2* system model

The M/M/1/2* queue can be described as a three-state Markov chain, where each state represents an empty system, a single packet receiving service or a system with two packets, one in service and one in the queue. Therefore, the steady state probabilities of the system are the same with the M/M/1/2 model. Moreover, the number of packets in the system is not affected by the replacements taking place, thus, the probabilities $\mathbb{P}(\psi_i)$ and $\mathbb{P}(\bar{\psi}_i)$ are given by (3.71) and (3.72), respectively. The average number of packets in the system is $\mathbb{E}[N] = 0p_0 + 1p_1 + 2p_2$, and the effective arrival rate is $\lambda_e = (1/\mathbb{E}[Y])(1-p_2)$, as in the M/M/1/2 model. The second term in (3.52) $\frac{1}{2}\mathbb{E}[Z_i^2]$ also remains the same as obtained in (3.73), since conditioning on the events ψ_i and $\bar{\psi}_i$ the distribution of the interdeparture time is independent of the system time and the selected queuing model.

To characterize the average age of the system, we need to calculate the terms $\mathbb{E}[T_{i-1}|\psi_{i-1}]$ and $\mathbb{E}[T_{i-1}|\bar{\psi}_{i-1}]$. In the M/M/1/2* queue the packets waiting for transmission are replaced by newly generated packets. In this case, the age is updated by the packets that are served. A packet may spent some time waiting in the queue and then be discarded by a newly received packet or entering the server. In order to characterize the system time for transmitted packets let T be the system time for any packet, which may or may not be discarded. Then, the probability that the system time surpasses a threshold can be partitioned in the following three events; a packet either receives service if the server is idle, or waits and then is transmitted (tx) if the server is busy, or gets dropped ($drop$). Thus, we have

$$\mathbb{P}(T>t) = \mathbb{P}(\text{idle})\mathbb{P}(T>t|\text{idle}) + \mathbb{P}(\text{busy, tx})\mathbb{P}(T>t|\text{busy, tx}) + \mathbb{P}(\text{busy, drop})\mathbb{P}(T>t|\text{busy, drop}). \tag{3.81}$$

The average age for the M/M/1/2* model is calculated similarly to the M/M/1/2 model. The resulting conditional expectations of the system time T_{i-1} are

$$\mathbb{E}[T_{i-1}|\psi_{i-1}] = \frac{\lambda}{(\lambda+\mu)^2} + \frac{1}{(\lambda+\mu)} = \frac{2\lambda+\mu}{(\lambda+\mu)^2}, \tag{3.82}$$

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

$$\mathbb{E}[T_{i-1}|\bar{\psi}_{i-1}] = \frac{\lambda}{(\lambda+\mu)^2} + \frac{1}{\mu} + \frac{1}{(\lambda+\mu)} = \frac{1}{\mu} + \frac{2\lambda+\mu}{(\lambda+\mu)^2}. \tag{3.83}$$

Substituting the probabilities (3.71), (3.72), the conditional expectations of the system time (3.82), (3.83) and the conditional expectations of the interdeparture time (3.56), (3.59), to (3.61) we obtain

$$\mathbb{E}[T_{i-1}Z_i] = \frac{1}{\mu^2} + \frac{1}{\lambda\mu} - \frac{2\lambda+\mu}{(\lambda+\mu)^3}. \tag{3.84}$$

Finally, using (3.73) and (3.84) the average Age of Information is calculated as

$$\Delta_{\mathrm{M/M/1/2^*}} = \frac{1}{\lambda} + \frac{2}{\mu} + \frac{\lambda}{(\lambda+\mu)^2} + \frac{1}{\lambda+\mu} - \frac{2(\lambda+\mu)}{\lambda^2+\lambda\mu+\mu^2}. \tag{3.85}$$

For infinitely large arrival rate λ the average age yields

$$\lim_{\lambda\to\infty} \Delta_{\mathrm{M/M/1/2^*}} = \frac{2}{\mu}. \tag{3.86}$$

This value is equal to the just-in-time lower bound for a queue with an FCFS discipline presented in Section 2.2. Intuitively, for very high arrival rates, packets containing fresh information are always available for transmission in both the M/M/1/1 and the M/M/1/2* queues.

For the peak age, the probability that it surpasses a threshold is

$$\begin{aligned}\mathbb{P}(A>\alpha)_{\mathrm{M/M/1/2^*}} &= \frac{e^{-(\lambda+\mu)\alpha}}{\lambda(\lambda+\mu)(\lambda-\mu)}(\lambda^3 - 3\mu^3 + \lambda\mu(\lambda+\mu)\\ &(1+(\lambda-\mu))) + \frac{e^{-\mu\alpha}}{\lambda(\lambda+\mu)(\lambda-\mu)}(3\mu^3 + \lambda(\lambda+\mu)(\lambda-\mu)\\ &+ \lambda\mu\alpha(\lambda+\lambda\mu-2\mu^2)) - \frac{e^{-\lambda\alpha}}{(\lambda+\mu)(\lambda-\mu)}(\lambda^2-\lambda\mu+\mu^2).\end{aligned} \tag{3.87}$$

Finally, the average peak age can be calculated directly as

$$A_{\mathrm{M/M/1/2^*}} = \mathbb{E}[Z+T] = \frac{1}{\mu} + \frac{\lambda}{(\lambda+\mu)^2} + \frac{1}{\lambda} + \frac{1}{\mu}\frac{1}{(\lambda+\mu)}. \tag{3.88}$$

The M/M/1/(N+1)* system model

A fourth packet management policy can be viewed as an extension of the M/M/1/2* model for the system of Figure 3.1 consisting of N independent sources. This model will be identified as M/M/1/(N+1)*. The M/M/1/(N+1)* queue keeps at most N packets waiting for transmission, one for each source. The packets waiting for transmission are replaced by newly generated packets if any. Upon arrival of a new status update, we examine the queue

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

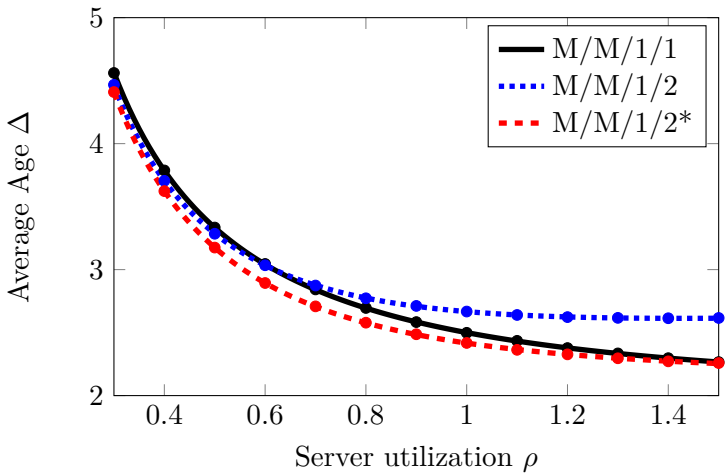

Figure 3.11: Average age vs server utilization of the M/M/1/1, M/M/1/2, and M/M/1/2* queues, for $\mu = 1$ [12].

for an existing packet of the same source. If such a packet exists it is replaced with the new arrival. Otherwise, the packet is placed at the end of the queue. The queue evolution can be modeled as a discrete time Markov chain, where each state represents the packets in the system. For more details we refer the reader to [45].

Numerical results

In [12] the performance of the three packet management schemes has been evaluated and compared with some of the models discussed previously. Performance is measured with respect to the Age of Information and the Peak Age of Information. For service rate $\mu = 1$ we examine the average behavior of our performance metrics as a function of the arrival rate λ.

In Figure 3.11, we plot the average AoI for the M/M/1/1, M/M/1/2, and M/M/1/2* models versus ρ. The average age is monotonically decreasing for all the three schemes, and the M/M/1/2* model outperforms the other two for the entire range $0 < \rho \leq 1.5$. When $\lambda = 0.6$, the average age of the M/M/1/2* case is approximately 5% smaller compared to the M/M/1/1 and M/M/1/2. At $\lambda = 0.6$ there is also a cross point where the $\Delta_{\mathrm{M/M/1/1}}$ curve achieves smaller values than the $\Delta_{\mathrm{M/M/1/2}}$ curve. This is due to the fact that as the arrival rate increases the packet waiting in the queue becomes less useful and the absence of buffering becomes a preferable option.

Figure 3.12 illustrates the average age for the M/M/1 using FCFS, M/M/1/2*, M/M/1 using LCFS with preemption, and M/M/∞ models. In the first model a packet is stored in the queue if the server is busy and the next packet that receives service is the one at the

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

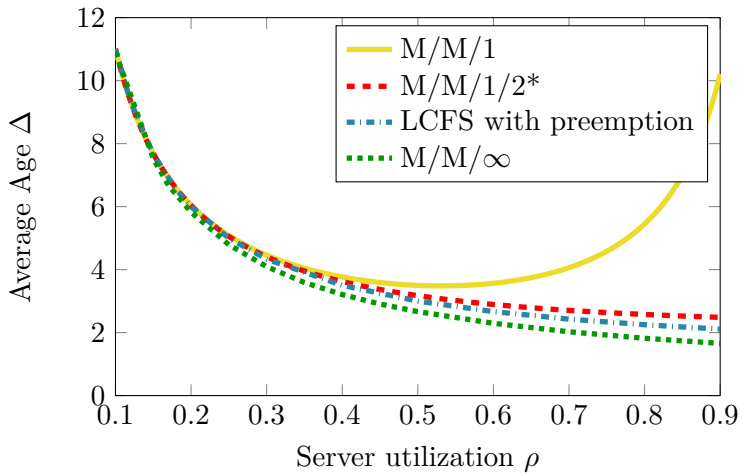

Figure 3.12: Average age vs server utilization of the M/M/1, M/M/1/2*, and M/M/∞ queues, for $\mu = 1$.

head of the queue. In the second model, at most two packets can be kept in the system; one in the server and one in the queue, and the packet waiting for transmission can be replaced with a newly arrived packet if any. In the third model, a packet arrival preempts the packet currently in service, if any. In the fourth model we consider an infinite number of servers which means that there is no queue and all packets are instantaneously served upon generation. For $\lambda \ll \mu$, the four models have similar behavior since the small arrival rate is the one that determines the age at the destination node, that is, the larger the interarrival time the smaller the effect of the respective model to the result. However, as the arrival rate increases and specifically for $\rho > 0.5$, we observe that the $\Delta_{\mathrm{M/M/1}}$ curve increases with ρ due to backlog that causes queuing delay. Recall that for the M/M/1 the age approaches infinity as ρ approaches 1. This effect is eliminated with the rest three schemes which provide significant improvements in the average age. As ρ increases, the gap between the $\Delta_{\mathrm{M/M/1/2^*}}$ and the $\Delta_{\mathrm{M/M/\infty}}$ curve increases as well. As $\lambda \to \infty$ the average age approaches $1/\mu = 1$ when preemption is allowed. This corresponds to a 50% reduction in minimum age compared to the just-in-time lower bound. The M/M/∞ model serves every new arrival immediately, hence more frequent transmissions lead to a more updated destination node. However, this comes with the cost of wasted network resources due to packets rendered obsolete as discussed in Section 3.4.1. Indicatively, for $\lambda = 1.5$ the performance of the M/M/∞ model is approximately 53% better that the M/M/1/2*.

To illustrate the behavior of the packet management schemes with respect to the peak age, Figure 3.13 shows the complementary cumulative distribution function for the peak age obtained by the equations (3.68), (3.79) and (3.87). We consider service rate $\mu = 1$, and the arrival rates $\lambda = 0.5$ and $\lambda = 1.3$. For all the three models, the probability that the

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

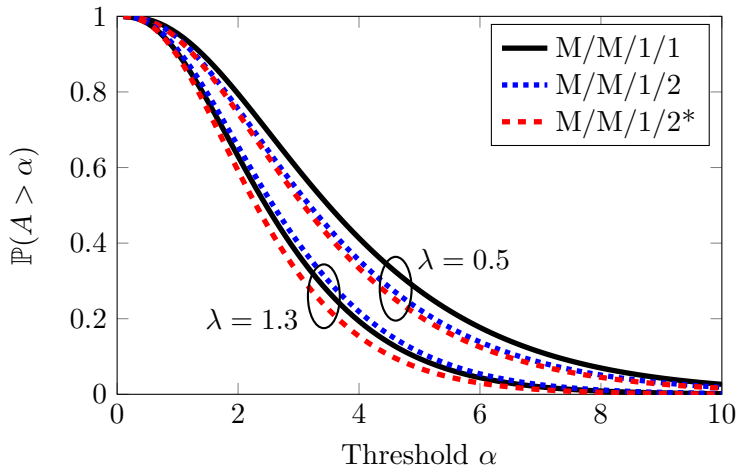

Figure 3.13: Probability that the peak age surpasses a threshold α vs threshold α for the M/M/1/1, M/M/1/2, and M/M/1/2* queues, for $\lambda = 0.5$ and $\lambda = 1.3$ [12].

peak age surpasses a threshold α reduces for a larger arrival rate λ. The M/M/1/2* queue results in smaller peak age over the entire threshold range. The M/M/1/2 model performs better than the M/M/1/1 model for smaller arrival rates such us $\lambda = 0.5$, but has worse performance for larger arrival rates such as $\lambda = 1.3$.

3.6 Summary

Continuing on the queue-theoretic model of age in this section we provided a host of results for different queue disciplines and packet management techniques, treating age as a metric to be optimized. Through these early works, the need for metrics based on age capturing timeliness in a more tractable fashion became evident, and this led to the introduction of Peak Age of Information (PAoI) which was also presented here. This understanding has also led to spawn more metrics relevant to AoI, such as the Age Penalty, and the similar Cost of Update Delay but also the Value of Information of Update which we will see in later sections. In what follows we discuss how age can be minimized introducing rate control or introducing packet deadlines, and finally how age can be jointly optimized with throughput and delay.

4 Subsequent Works and Extensions

The previous section illustrated the potential of the AoI metric in a large scope of systems. How to minimize the AoI at a destination depends on the topology, the characteristics of

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

the network, and the available means to control it. There is a growing body of literature that recognises this fertile ground and tries to extend the early works in the field either by introducing new models of the medium or by controlling different parts of the system.

In this section, we first assume an a-priori knowledge of the server state and look at the optimal rate control policy with respect to the freshness of the information at the destination. In addition, we present a packet management mechanism that controls packets by attaching a deadline constraint to them. It will be shown that the packet deadline in a system with capacity two can outperform both the M/M/1/1 and M/M/1/2 systems without a control mechanism. Furthermore, we present an optimization problem with different objectives such as age, peak age, throughput or delay, and efficient algorithms to find the optimal policies are discussed.

4.1 Rate Control Optimization

An approach to manage the timeliness of the transmitted information from a source to a remote destination, is to assume perfect knowledge of the service state. With such a knowledge, updates can be generated only when the server is not busy, as shown in Figure 4.1. Thus, the waiting time W at the queue is eliminated and the system time T consists only of the service time S. In contrast to the models presented until now, we assume that the source is able to generate status updates on demand, thus, we control the Age of Information of the system by considering the sampling to be a function of the service process.

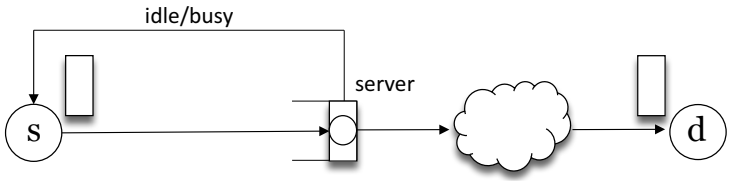

Figure 4.1: System model with feedback.

The *just-in-time* policy, meaning that, update i is generated as soon as update $i-1$ finishes service, might seem the optimal policy. This policy was defined in [52] as the *zero-wait* policy. Here, we illustrate that the zero-wait policy is not necessarily optimal with respect to the average Age of Information of the system. This policy achieves the maximum throughput and the minimum delay but is not the optimal with respect to AoI. Moreover, we investigate the conditions under which the zero-wait policy is optimal.

In Figure 4.2, an example of the age process of the presented system model is depicted. Once the service of update i is finished the source may wait a time interval $L_i \in [0, M]$ before extracting update $i+1$ from the stochastic process under observation $X(t)$. The

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

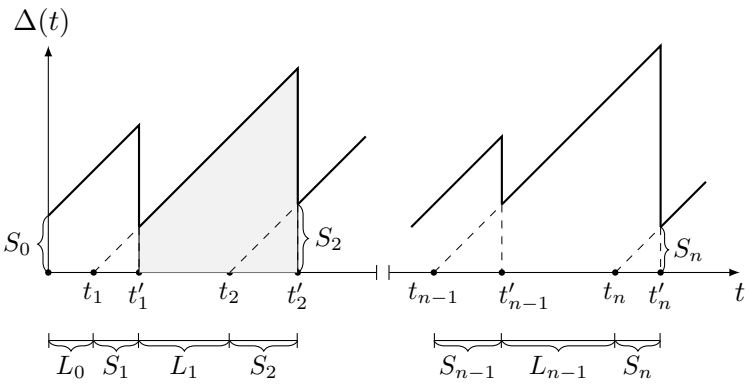

Figure 4.2: Example of age evolution for the system model in Figure 4.1 [52].

variable M denotes the maximum waiting period allowed by the system. The service time of update i is denoted by S_i. The service process $(S_0, S_1, \dots)$ is modeled as a stationary ergodic Markov chain with an uncountable state space and positive and finite mean $\mathbb{E}[S_i]$. The proposed modeling eliminates the memoryless property of the system and generalizes the i.i.d. assumption of the service times holding in the previous sections. Later we will see that the zero-wait policy is close to optimum when the service times are highly random.

Consider the definition of the time average AoI of a status update system in (2.7), and the decomposition into disjoint geometric parts in (2.8). The areas Q_i are calculated as

$$Q_i = \int_{t_{i-1}}^{t'_i} (t - t_{i-1})\,\mathrm{d}t - \int_{t_i}^{t'_i} (t - t_i)\,\mathrm{d}t \overset{(a)}{=} \int_{S_i}^{S_i+L_i+S_{i+1}} t\,\mathrm{d}t, \tag{4.1}$$

where equality (a) follows from the fact that the time interval (t'_{i-1}, t_i) is well defined in contrast with the basic system model of Figure 2.2. We rewrite the disjoint parts as $Q_i = q(S_i, L_i, S_{i+1})$ where

$$q(s, l, s') = \int_{s}^{s+l+s'} t\,\mathrm{d}t. \tag{4.2}$$

The objective of the status update system is to minimize the average Age of Information by controlling the sequence of waiting times $(L_0, L_1, \dots)$. Let $\pi \doteq (L_0, L_1, \dots)$ denote a status update policy. Let Π denote the set of all causally feasible policies satisfying $L_i \in [0, M]$ for all i. The time interval L_i is determined based on past realizations of the service process $(S_0, S_1, \dots, S_i)$. The conditional distribution of $(S_{i+1}, S_{i+2}, \dots)$ based on $(S_0, S_1, \dots, S_i)$ is available. Then, the stochastic optimization problem of minimizing the

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

average Age of Information can be formulated as [52]

$$\begin{aligned} c_{\text{opt}} = \min_{\pi\in\Pi} \quad & \lim_{n\to\infty} \sup \frac{\mathbb{E}[\sum_{i=0}^{n-1} q(S_i, L_i, S_{i+1})]}{\mathbb{E}[\sum_{i=0}^{n-1}(S_i + L_i)]} \\ \text{s.t.} \quad & \lim_{n\to\infty} \inf \frac{1}{n}\mathbb{E}\left[\sum_{i=0}^{n-1}(S_i + L_i)\right] \geq T_{\min}, \end{aligned} \tag{4.3}$$

where the expectation is taken over the service process $(S_0, S_1, \dots)$ for a given policy π; $T_{\min}$ is the minimum average waiting time of the source due to hardware and physical constraints, hence M should exceed $T_{\min}$. Note that $(S_0, S_1, \dots)$ is stationary and ergodic thus $\mathbb{E}[\sum_{i=0}^{n-1}(L_i + S_{i+1})] = \mathbb{E}[\sum_{i=0}^{n-1}(S_i + L_i)]$.

The problem in (4.3) belongs to a class of constrained semi-Markov decision processes (SMDP) with a possibly uncountable state space, which is well-known for its difficulty. The authors in [52] prove that there exists a *stationary randomized* policy that is optimal for the problem in (4.3). Further, it is proven that there is an optimal *stationary deterministic* policy for this problem. Finally, a low-complexity algorithm is developed to find the optimal stationary deterministic policy that solves problem (4.3).

Consider two possible scenarios regarding the service time S. If $T_{\min} \leq \mathbb{E}[S]$, the zero-wait policy is the preferable choice, i.e. $\pi_{\text{zero-wait}} = (0, 0, \dots)$. If $T_{\min} > \mathbb{E}[S]$, the minimum average waiting time is $\mathbb{E}[l(S)] = T_{\min} - \mathbb{E}[S]$, but the optimal policy may also add waiting time such that $\mathbb{E}[l(S)] > T_{\min} - \mathbb{E}[S]$.

The solution to problem (4.3) can be generalised to any measurable, non-negative, monotonically increasing function of time. Considering age as a special case of such a function and i.i.d. service times S_i, leads to a simple solution that explicitly indicates whether the optimal policy is the minimum wait or not. The trapezoid areas shown in Figure 4.2 can be calculated as

$$q(s, l, s') = \frac{1}{2}(2s + l + s')(l + s'). \tag{4.4}$$

Then, the expectation of $q(s, l, s')$ is given by

$$\begin{aligned} \mathbb{E}[q(S, l(S), S')] &= \mathbb{E}[\frac{1}{2}(S + l(S))^2 + (S + l(S))S'] \\ &= \frac{1}{2}\mathbb{E}[(S + l(S))^2] + \mathbb{E}[(S + l(S))]\,\mathbb{E}[S']. \end{aligned} \tag{4.5}$$

Then (4.3) can be reformulated as

$$c_{\text{opt}} = \min_{l\in L^2(\mu_S)} \quad \frac{\mathbb{E}[(S + l(S))^2]}{2\mathbb{E}[(S + l(S))]} + \mathbb{E}[S] \tag{4.6}$$

$$\text{s.t.} \quad \mathbb{E}\left[(S + l(S))\right] \geq T_{\min} \tag{4.7}$$

$$0 \leq l(s) \leq M,\ \forall s \geq 0,$$

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

where μ_S is the probability measure of S_i and the function $l:\ [0,\infty) \to [0,M]$ belongs to the Lebesgue space $L^2(\mu_S)$ [47, Ch. 3], because

$$\int_0^\infty |l(s)|^2 \, \mathrm{d}\mu_S(s) \leq \int_0^\infty M^2 \, \mathrm{d}\mu_S(s) = M^2 < \infty. \tag{4.8}$$

Lemma 4.1.1. The functional $h : L^2(\mu_S) \to \mathbb{R}$ defined by

$$h(l) = \frac{\mathbb{E}[(S+l(S))^2]}{\mathbb{E}[S+l(S)]} \tag{4.9}$$

is convex on the domain

$$\mathbf{dom}\, h = \{l : l(s) \in [0,M], \forall s \geq 0,\ l \in L^2(\mu_S)\}. \tag{4.10}$$

Proof. See [52] and references therein. □

The Karush-Kuhn-Tucker (KKT) theorem for infinite dimensional space and the calculus of variations are used to obtain the following result.

Theorem 4.1.2 (Optimal solution of β minimum policy)**.** The optimal solution to problem (4.6) is

$$l(s) = (\beta - s)_0^M, \tag{4.11}$$

where $(x)_0^M = \min[\max[x,0],M]$ and $\beta > 0$ satisfies

$$\mathbb{E}[S+l(S)] = \max\left(T_{\min}, \frac{\mathbb{E}[(S+l(S))^2]}{2\beta}\right). \tag{4.12}$$

Proof. See [52] and references therein. □

Equation (4.11) implies $(\beta)_S^{M+S} = S + l(S)$. If $T_{\min} \geq \frac{\mathbb{E}[(S+l(S))^2]}{2\beta}$, then the optimal policy $l(\cdot)$ satisfies

$$\mathbb{E}[S+l(S)] = T_{\min} \geq \frac{\mathbb{E}[(S+l(S))^2]}{2\beta}, \tag{4.13}$$

and the minimum average waiting time is $\mathbb{E}[l(S)] = T_{\min} - \mathbb{E}[S]$. If $T_{\min} < \frac{\mathbb{E}[(S+l(S))^2]}{2\beta}$, then the optimal policy $l(\cdot)$ satisfies

$$\mathbb{E}[S+l(S)] = \frac{\mathbb{E}[(S+l(S))^2]}{2\beta} > T_{\min}, \tag{4.14}$$

and the average waiting time is $\mathbb{E}[l(S)] > T_{\min} - \mathbb{E}[S]$. For the special case $T_{\min} = 0$, constraint (4.7) is always satisfied and we have $T_{\min} \leq \mathbb{E}[S]$. By substituting $T_{\min} = \mathbb{E}[S]$ into (4.13) and (4.14) we obtain the criterion on whether the zero waiting policy is optimal.

Finally, we compare the optimal policy provided by Theorem 4.2 with two "default" policies for two models of the service processes. The substance of the result does not depend on the distribution of the service time S_i but the performance depends on the properties of the service process. The two policies are the following:

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

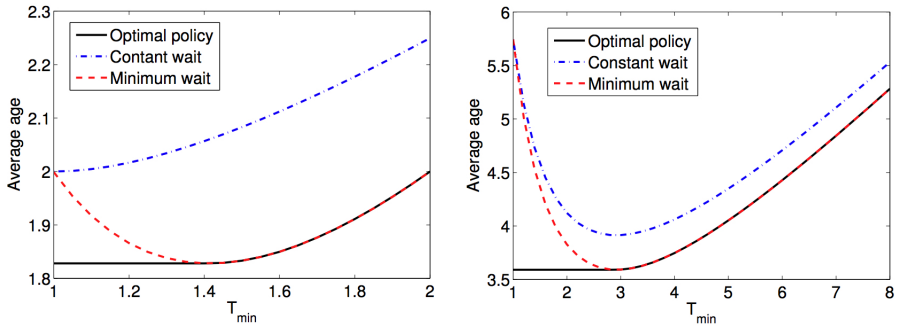

Figure 4.3: Average age vs $T_{\min}$ for i.i.d. (a) discrete service times and (b) log-normal distributed service times [52].

- *Constant wait*: Each update is followed by a constant delay L before submitting the next update with $const = T_{\min} - \mathbb{E}[S]$.
- *Minimum wait*: An update is followed by a delay given by the deterministic function $L = l(S_i)$, where $l(\cdot)$ is given by (4.11) and β is chosen to satisfy $\mathbb{E}[l(S)] = T_{\min} - \mathbb{E}[S]$.

When $T_{\min} = \mathbb{E}[S]$, both constant wait and minimum wait policies reduce to the zero wait policy. Figure 4.3 shows the average Age of Information of the system as a function of $T_{\min}$ for i.i.d. (a) discrete and (b) log-normal distributed service times. In both plots, the constant wait policy results in the largest AoI over all $T_{\min}$. When $T_{\min}$ surpassed a certain threshold the optimal policy meets the constraint (4.7) with equality. For small values of $T_{\min}$ up to that threshold, the constraint is not active and the minimum wait policy deviates from the optimum. When $T_{\min}$ surpassed that threshold the optimal policy meets the constraint (4.7) with equality. Looking deeper into general functions and correlated service processes, it can be shown that the zero-wait policy is optimal if the correlation coefficient between S_i and S_{i+1} is -1, or the service times are equal to a constant value or age is a constant function.

4.2 Packet Deadlines

Another interesting concept in the research of the AoI is that of packet deadlines [28]. Instead of using different queueing models to optimize AoI, a deadline is imposed on packets waiting in the queue and they are dropped in case they are still waiting for transmission once their deadline expires. Intuitively, longer deadlines lead to stale updates occupying the system and unnecessarily outdated status information potentially arriving at the destination, while a shorter deadline would mean that fresh updates are discarded without updating the destination and reducing age.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

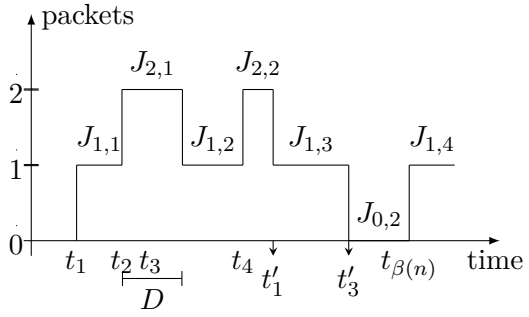

Figure 4.4: Example of queue evolution on time spent in each state.

For the analysis an M/M/1/2 queue is considered that can be described as a three-state Markov chain. Each state of the Markov chain represents an empty system, a single packet receiving service or a system with two packets, one in service and one in the queue. However, assuming fixed deadline we lose the memoryless property of the system. Furthermore, the steady state probabilities of the system depend on the deadline period and not just the arrival and service times as in Subsection 3.5.

To derive the average Age of Information of the system, (3.52) is used. Recall that the random variables Z_i and T_i, denote the interdeparture and system time of update i, respectively, refer to transmitted packets. Then, we need to calculate the effective arrival rate λ_e and the $\mathbb{E}[ZT]$ and $\mathbb{E}[Z^2]$. Before calculating these terms we present some preliminary results of the queue in equilibrium.

Let p_0, p_1, and p_2 be the steady state probabilities of the number of packets in the system. Since classic results from queueing theory cannot be applied, a sample average that converges to the stochastic average is used as follows. Let J_{ij} be the duration of the jth visit to the state with i packets, as shown in Figure 4.4. Let $\alpha(\mathcal{T})$, $\beta(\mathcal{T})$, and $\gamma(\mathcal{T})$ be the total number of visits to state 0, 1, and 2, respectively. Suppose that our interval of observation is $(0, \mathcal{T})$. Then, the percentage of time spent in state 0, as $\mathcal{T} \to \infty$, is

$$\begin{aligned} p_0 &= \lim_{\mathcal{T}\to\infty} \frac{\sum_{j=1}^{\alpha(\mathcal{T})} J_{0j}}{\mathcal{T}} \\ &= \lim_{\mathcal{T}\to\infty} \frac{\sum_{j=1}^{\alpha(\mathcal{T})} J_{0j}}{\sum_{j=1}^{\alpha(\mathcal{T})} J_{0j} + \sum_{j=1}^{\beta(\mathcal{T})} J_{1j} + \sum_{j=1}^{\gamma(\mathcal{T})} J_{2j}}. \end{aligned} \tag{4.15}$$

After calculations we have

$$p_0 = \frac{\pi_0 \mathbb{E}[J_0]}{\pi_0 \mathbb{E}[J_0] + \pi_1 \mathbb{E}[J_1] + \pi_2 \mathbb{E}[J_2]}. \tag{4.16}$$

where, π_i is the percentage of visits to state i and $\mathbb{E}[J_i]$ is the average time spent in state i. The expressions for the probabilities p_1 and p_2 are derived in the same manner. Then,

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

applying a Markov chain analysis to the process of visits to state i, we obtain the equilibrium distribution of visits

$$\pi_0 = \frac{\mu}{2(\lambda+\mu)}, \quad \pi_1 = \frac{1}{2}, \qquad \pi_2 = \frac{\lambda}{2(\lambda+\mu)}. \tag{4.17}$$

The Markov chain is shown in Figure 4.5.

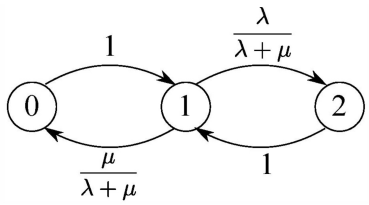

Figure 4.5: The Markov chain of the number of packets in the system [28].

Next, we present the derivation of the average time spent in each state. The time spent in state 0 is equal to the interarrival time, i.e.,

$$\mathbb{E}[J_0] = \frac{1}{\lambda}. \tag{4.18}$$

The time spent in state 1 is equal to

$$\begin{aligned}\mathbb{E}[J_1] &= \mathbb{P}(Y < S)\mathbb{E}[Y|Y < S] + \mathbb{P}(Y > S)\mathbb{E}[Y|Y > S] \\ &= \frac{\lambda}{\lambda+\mu}\frac{1}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}\frac{1}{\lambda+\mu} = \frac{1}{\lambda+\mu}.\end{aligned} \tag{4.19}$$

Packets are dropped from the system whenever the server is busy for a period longer than the deadline, thus the time spent in state 2 is

$$\mathbb{E}[J_2] = \frac{1}{\mu}(1 - e^{-\mu D}). \tag{4.20}$$

Finally, the percentage of the time spent in each state is obtained as

$$p_0 = \frac{\mu^2}{\mu^2 + \lambda\mu + \lambda^2(1 - e^{-\mu D})}, \tag{4.21}$$

$$p_1 = \frac{\lambda\mu}{\mu^2 + \lambda\mu + \lambda^2(1 - e^{-\mu D})}, \tag{4.22}$$

$$p_2 = \frac{\lambda^2(1 - e^{-\mu D})}{\mu^2 + \lambda\mu + \lambda^2(1 - e^{-\mu D})}. \tag{4.23}$$

Packets that receive service are those that don't find a full system and are not dropped due to the packet deadline. The effective arrival rate is $\lambda_e = \lambda(1 - p_1 e^{-\mu D} - p_2)$. To calculate

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

the first and second term in (3.52), we follow the approach in Subsection 3.5. That is, we describe the event ψ_{i-1} such that Z_i and T_{i-1} are conditionally independent. Conditioning on whether the $(i-1)$th packet leaves behind an empty system, we are able to calculate the second moment of the interdeparture time as follows

$$\begin{aligned}\frac{1}{2}\mathbb{E}[Z_i^2] &= \frac{1}{2}\left(\mathbb{P}(\psi_{i-1})\mathbb{E}[Z_i^2|\psi_{i-1}] + \mathbb{P}(\bar{\psi}_{i-1})\mathbb{E}[Z_i^2|\bar{\psi}_{i-1}]\right)\\ &= \frac{1}{2}\left(\frac{\mu}{\mu+\lambda(1-e^{-\mu D})}\frac{2(\lambda^2+\lambda\mu+\mu^2)}{(\lambda\mu)^2} + \left(1-\frac{\mu}{\mu+\lambda(1-e^{-\mu D})}\right)\frac{2}{\mu^2}\right)\\ &= \frac{(\mu(\lambda^2+\lambda\mu+\mu^2)+\lambda^3(1-e^{-\mu D}))}{\lambda^2\mu^2(\mu+\lambda(1-e^{-\mu D}))}. \end{aligned} \tag{4.24}$$

Next, the expectation $\mathbb{E}[T_{i-1}Z_i]$ will be calculated using the partition (3.61). To characterize the conditional distribution of the system time T_{i-1} given the events ψ_{i-1} and $\bar{\psi}_{i-1}$, we write the system time of update $(i-1)$ as $T_{i-1} = W_{i-1} + S_{i-1}$, where W_{i-1} is the waiting time and S_{i-1} is the service time of update $(i-1)$. The waiting time of a packet is zero, if the system is empty and it is equal to the service time of the previous packet if the server is busy. The remaining waiting time until the next departure is exponentially distributed with mean $1/\mu$. Note that the waiting time of packet $(i-1)$ is independent of future arrivals and therefore of the event ψ_{i-1}. Moreover, for the conditional expected service time we need to consider the state of the system as a function of time. We refer the reader to [28] for more details.

For infinitely large value of arrival rate λ the average age yields

$$\lim_{\lambda\to\infty}\Delta_{\mathrm{M/M/1/2}_D} = \frac{3}{\mu} - \left(D+\frac{1}{\mu}\right)e^{-\mu D}. \tag{4.25}$$

For infinitely large value of packet deadline D the average age yields

$$\lim_{D\to\infty}\lim_{\lambda\to\infty}\Delta_{\mathrm{M/M/1/2}_D} = \frac{3}{\mu}. \tag{4.26}$$

This value is equal to the average age for the M/M/1/2 queue for infinitely large arrival rate λ (3.78). Setting $D=0$ results in $\Delta_{\mathrm{M/M/1/2}_D} = \frac{2}{\mu}$ that is equal to the average age for the M/M/1/1 queue for infinitely large arrival rate λ (3.66).

Figure 4.6 shows the average Age of Information for the $\mathrm{M/M/1/2}_D$ queue as a function of the packet deadline D, for service rates $\mu=1$ and $\mu=2$. The minimum age for each arrival rate λ is indicated with $\triangle$. Observe that $\Delta_{\mathrm{M/M/1/2}_D}$ initially decreases for a small range of D and then increases with D toward an asymptotic value. For fixed D, as λ increases, the age decreases since the destination is updated more often. The results show that the deadline has a larger impact for higher λ. Furthermore, for $\mu=2$ the age is smaller than $\mu=1$, and the deadline effect is more pronounced.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

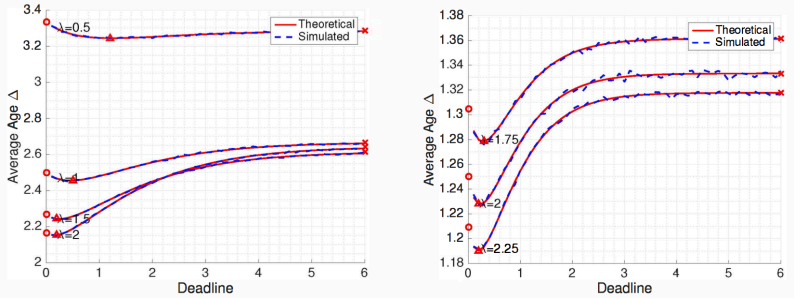

Figure 4.6: Average age vs deadline of the M/M/1/2$_D$ queue, for $\mu = 1$ (left) and $\mu = 2$ (right) and various arrival rates λ. With $\triangle$ are the minimum values, with $\circ$ the M/M/1/1 age and with $\times$ the M/M/1/2 age [28].

If $D = 0$, the system is equivalent to the M/M/1/1 queue where no packets are kept in the queue waiting for transmission. In addition, as the deadline goes to infinity the system become equivalent to the M/M/1/2 queue. Therefore, the packet deadline scheme can be seen as a way to transit between the M/M/1/1 and the M/M/1/2 queue. The results indicate that the performance can be actually improved.

4.3 Optimizing Age, Throughput, and Delay

In [4], an information-update system is considered, where updates in form of packets are stored in a queue and then are forwarded to a remote destination through multiple servers, as shown in Figure 4.7. In this model, the status updates do not necessarily arrive in the order of their generation times [4]. Two fundamental topics are addressed (i) Establishing age-optimality in a general policy space under an arbitrary arrival process and (ii) Simultaneously optimizing multiple performance metrics, such as age, throughput, and delay. Furthermore, one of the key outcomes of this work is that a setup with multiple servers outperforms in terms of age, the setup of one single server under the same policy.

The authors prove that if the packet service times are i.i.d. exponentially distributed, then for an arbitrary arrival process and any queue size, a preemptive last-generated first-served (LGFS) policy achieves an age process that is stochastically smaller than any causally feasible policy.

This implies that the preemptive LGFS policy minimizes many data freshness metric, including time-average age, average PAoI, and the time-average *age penalty*. The intuition is that the freshest update packets are served as early as possible in the preemptive LGFS policy. The distribution of the age process of the preemptive LGFS policy is invariant over all queue sizes.

[4] This is the first work that considered this case.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

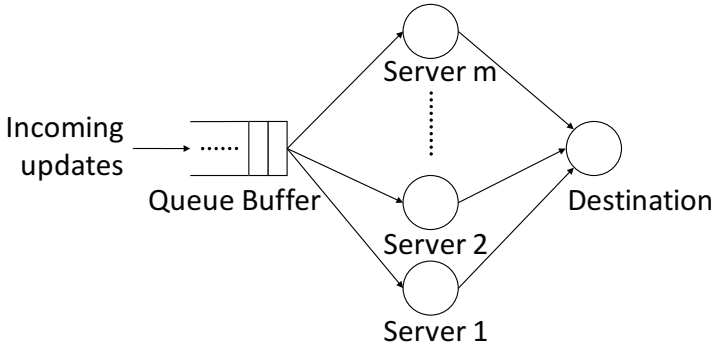

Figure 4.7: The system model of [4], comprising one source, m servers, and a remote destination.

Regarding the age penalty, introduced in [52], the authors define an age penalty function to model the level of "dissatisfaction" due to data staleness or the "need" for a new information update. This function is a general non-decreasing function, that depends on specific applications. Note that the Cost of Update Delay (CoUD) metric introduced in [36], which will be outlined as a recent work in Section 6, could serve as an alternative measure of age penalty.

If the packet service times are i.i.d. exponentially distributed across time and servers, then for all packet generation and arrival times, and for the case of infinite buffer size the preemptive LGFS policy is throughput-optimal and mean delay-optimal among all the causal policies.

The plots in Figure 4.8 present the time average age versus the server utilization ρ for the scenario where five servers are available. A first comment is that the age performance for each policy is better than the case with one server because of the diversity.

The preemptive LGFS policy achieves the best age performance among all the considered policies. As mentioned also earlier, the age performance of the preemptive LGFS policy is the same for any queue size B.

However, the age performance of the non-preemptive LGFS policy and the FCFS policy depends on the queue size B when there are multiple servers. Furthermore, in case of FCFS policy with $B = \infty$, the average age is increasing rapidly when the traffic intensity is high. The reason behind this is the increased congestion in the network leading to a delivery of stale packets. Moreover, in case of th e FCFS policy with $B = 10$, the average age is high but remains bounded at high traffic intensity. This is because a fresh packet has a better opportunity to be delivered in a relatively short period compared with FCFS policy with $B = \infty$.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

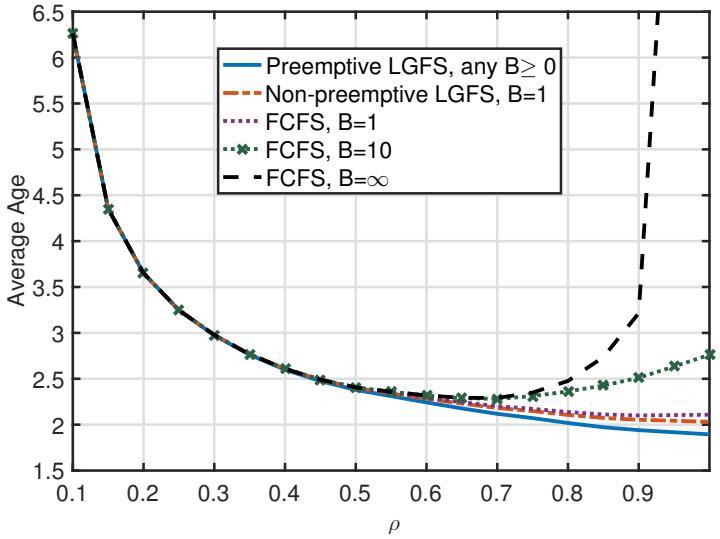

Figure 4.8: Average age vs server utilization ρ, for five servers with queue size B [4].

4.4 Summary

In this section we discussed how age can be minimized introducing rate control or introducing packet deadlines, and finally how age can be jointly optimized with throughput and delay. In what follows, the age metric is being treated as a tool to address different areas. Clearly the areas discussed are ones that have already begun to be explored.

5 Age as a Tool

This section departs from modeling and characterization of AoI and provides directions on how to apply the notion of AoI in different areas. Although the notion of AoI has only recently started evolving, there has already been a number of works that study its application as a tool. Clearly, this new concept can be applied towards achieving a wide range of objectives in communication systems that deal with time critical information while having limited resources. The range of possibilities may become more apparent to the reader in the next section where we will present the latest contributions in the field that cover an extended space.

5.1 Channel State Information

In cellular networks, each base station (BS) needs to estimate the channel responses from the user equipments (UEs) that are active in the current coherence block. The channel

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

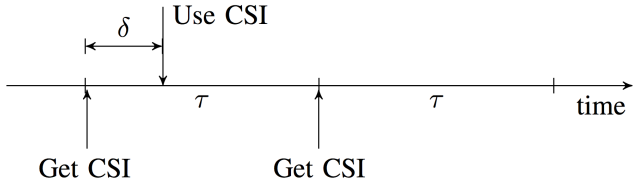

Figure 5.1: Periodic reports of CSI with period τ. The CSI is available after an interval δ [13], [14].

responses are utilized by the BS to process the uplink and downlink signals. In practice, these values need to be estimated regularly thus, they add a non negligible overhead to the system. The knowledge of the current channel response realizations is the *channel state information* (CSI). If the uplink and downlink are separated in frequency, for example using a frequency division duplex (FDD) protocol, i.e. the respective channels are different thus, we can not rely on reciprocity.

In non-reciprocal wireless links, the transmitter knows the current channel state through the CSI feedback sent from the receiver. In this case, the available information at the destination has aged over time, affecting the efficiency of the communication. The channel information ageing is caused by multiple factors such as (i) measuring times, (ii) transmission delay, (iii) processing time required to decode and estimate the channel quality, (iv) processing time to run adaptation functions, (v) frame times, and (vi) the interval between consecutive reports from the UE. In [13] and [14] the factors (i)-(v) are collectively represented by a random variable that captures the time elapsed from the generation up to the reception of the CSI. Factor (vi) is denoted by τ. The case of deterministic τ corresponds to a system with periodic feedback, while the case of random τ corresponds to a system with aperiodic feedback. In current standards such as Long Term Evolution (LTE) both feedback modes are supported [**Ch. 8**, 49]. Figure 5.1 illustrates the system with periodic feedback that is the subject of recent studies.

Applying this setup to the AoI framework discussed in the previous sections we are interested in the effect of outdated CSI on the performance of feedback links and protocols. In [13] and [14] utility functions are used as a general performance metric that accounts for various scenarios and includes the cost of feedback. Thus, the tradeoff between utility and frequency of reports can be characterized. As a first step, the channel in these works is modeled as a Finite State Markov Channel (FSMC) with two states representing the fading conditions, one *good* and one *bad*, yielding tractable analytic results.

Details regarding the probability of error on channel estimation and the proposed utility function can be found in [13] and [14]. For very large values of the feedback cost the utility value decreases as the frame duration τ increases. For very small values of the feedback cost, decreasing τ leads to a decrease in the utility value. Otherwise, we should expect an optimal operation point with respect to τ. In general, the higher the cost, the smaller the

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

utility value. Results on Rayleigh fading channels indicate that the cost-utility tradeoff is more prominent for smaller values of Doppler shift, and the degradation of the utility value with the CSI age is more severe for high values of Doppler shift.

CSI for wireless networks with reciprocal channels

The work in [34] considers theoretical bounds and protocols to estimate and disseminate the global CSI in wireless networks with reciprocal channels. The networks that are considered in that work are assumed to be fully connected. The term global CSI is to capture that each node maintains its own table of estimates for all the channels in the network, for example in an N-node wireless network there are $L = N(N-1)/2$ channels. The nodes obtain estimations of the channels to which they are not directly connected via CSI dissemination. Nodes disseminate CSI by including one or more CSI estimations in each packet transmission. Thus, the other nodes can learn the states of channels to which they are not directly connected. Each node in the network directly estimates the $N-1$ channels to which it is directly connected and estimates the remaining $L-N+1$ channels in the network by collecting the disseminated CSI. The global CSI can be utilized by the nodes in the network in order to allow them to adapt their roles or their operations dynamically over time, such as serving as a relay at one time point. In [34] the case that CSI errors are caused by time-variation and staleness is considered, i.e., the time elapsed from when the time-varying channel was estimated and the current time. In general, there is a deterioration of performance with respect to perfect CSI knowledge as CSI is ageing. The authors in that paper developed a general framework for quantifying the staleness of global CSI in wireless networks. The developed framework departs from the queueing theoretic analysis that is dominant, so far, in the works considering AoI. Bounds on the minimum achievable CSI staleness throughout a network are provided. In addition, CSI dissemination protocols are developed, for which performance evaluation in terms of the maximum and average staleness is provided. Since each node maintains its own global CSI and fully connected networks are considered, the total number of estimates scales as $O(N^3)$, thus, global CSI can be feasible for reasonably sized networks.

Furthermore, it is considered that each packet consists of the overhead, the data, and the disseminated CSI information. Denoting by D the number of words that describe the data and the overhead, and M the number of channel estimates, then the total packet length is $P = M + D$ words. The term word is a general time unit representing the amount of time required to transmit a single CSI estimate. The authors proved that the maximum staleness is lower bounded by $L(D+1)$ for the case where only one CSI estimate is disseminated by each packet, i.e. $M = 1$. Similarly, a lower bound on the average staleness is given by $\frac{N^3-3N^2+8N-8}{4N}(D+1)$. We refer the reader to [34] for more details and general results.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

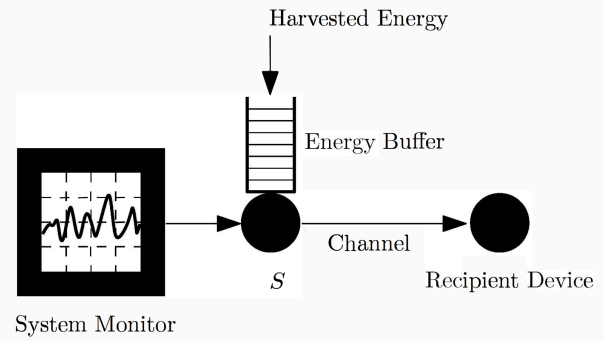

Figure 5.2: System model with an energy harvesting source [2].

5.2 Energy Harvesting

Consider the system model in Figure 5.2, consisting of a source - destination communication system with energy harvesting capability at the source. The time-varying availability of energy and battery constraints at the transmitter can limit the sampling rate at the source. Therefore, it is of interest to investigate how a stochastic energy harvesting system affects the AoI at the destination and to find the optimal policy.

The minimization of AoI under such considerations was studied in [2]. Consider the energy harvesting process $H(t)$ as the total energy harvested over time $[0, t]$. The number of status updates that can be served depends on $H(t)$. In the continuous time setup of the problem, the sequence of the extracted samples (i.e. status updates) should be selected to minimize the energy-constrained average AoI. The discrete time setup of the problem accounts for the state of the source that is either busy or idle, hence we have a binary decision problem at each time slot. Apart from the stochastic energy arrival process, another difference of this work compared to previous studies is the assumption of instant transmission of packets. A problem formulation that encounters transmission delays still needs to be determined. In this setup, the authors derive an offline solution that minimizes the time average AoI for an arbitrary energy replenishment profile, using a discrete time dynamic programming formulation. It is also found that the expected value of the current age as well as the current energy level at the transmitter is sufficient information to generate an optimal threshold policy. An effective online heuristic, named Balance Updating (BU), achieving performance close to an offline policy is also proposed. Simulations of the policies indicate that they can significantly improve the age over greedy approaches. Finally, the authors consider an extension of the previous formulation for stochastically generated updates.

Energy harvesting constraints are also considered in [56], where in contrast to [2] the randomness is on the service times and not on the energy arrival process. A complete representation of an energy harvesting system with a minimum AoI objective would include

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

the age process $\Delta(t)$, the available energy at the source $H(t)-U(t)$, where $U(t)$ is the energy consumption process, and the energy harvesting process $H(t)$. An optimal rate control policy would use this information to make the optimal decisions on the interarrival times of status updates. Such policies can be complex, even under simple harvesting assumptions. However, under the ergodicity assumption for the energy harvesting process,

$$\lim_{t\to\infty} H(t)/t = \eta, \tag{5.1}$$

follows a simpler approach that uses averaging [56]. From (5.1) and the arrival rate in (2.10) the condition $\lambda < \eta$ is obtained. Managing the average energy consumption is sufficient since in a sufficiently large battery removes the need for precise energy management. Then, the optimal rate control policy of this work assuming knowledge of the server state has been generalized and presented in Section 4.1. Parameter β of Theorem 4.2 can be chosen such that $\lambda = \eta - \epsilon$, for arbitrarily small $\epsilon > 0$.

The following two works have been presented in the time this volume was being put together and are thus, only briefly outlined.

First, [3] addresses the continuous time problem of optimizing when status updates should take place such that the average Age of Information will be minimized for a given energy harvesting profile. The online problem, where the energy profile is a stochastic process with known statistics is also investigated within a dynamic programming formulation. The key outcome is that tracking the expected value of the current age (a linear operation), together with the knowledge of the current energy level at the sender side is sufficient for generating an optimal threshold-type policy.

Second, [55], investigates a scenario where a sensor continuously monitors a system and transmits status updates to a destination. The sensor stores the harvested energy in a battery, and three cases are considered regarding the battery size, infinite, finite, and battery of size one. The authors design an optimal status update policy that minimize the long term average AoI. For the infinite battery scenario, a best-effort uniform status update policy is shown to be optimal. For the finite battery scenario, an energy-aware adaptive status update policy is asymptotically optimal, when the battery size goes to infinity. When the battery size is one, a threshold based status update policy is proposed and it is proved to be optimal.

In Section 6.2, we also outline work [1] that treats energy harvesting, where the communication model departs from single-hop.

5.3 Scheduling

Link scheduling is a fundamental problem in wireless communications. The problem has been widely studied in the context of access coordination. A subproblem of it, the so called

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

link activation problem, aims to determine links that can be simultaneously active in a shared channel. The optimal scheduling policy is usually governed by an objective related to some cost criterion. The most common performance objectives have been those of the maximum throughput or the minimum completion time for all the links.

The notion of AoI opens a new research direction for optimal scheduling, in particular because minimization of the completion time leads in general to suboptimal solutions in terms of AoI. The authors of [18] propose a novel approach of optimizing the transmission scheduling with respect to age. The system model consists of a set of source-destination pairs, or links, that share a common medium. Given the sets of packets residing at the sources, the task is to solve the minimum age scheduling problem (MASP), for which a schedule consists of activation of link subsets and the sequence of the activations. In addition to proving the NP-hardness of MASP, an integer linear programming (ILP) is provided for performance benchmarking. Moreover, a suboptimal but fast algorithm is provided. A similar problem is studied in [17]. There, the objective is to minimize the maximum Peak Age of Information, to address both information freshness and fairness among the sources.

As with the case for energy harvesting, works that have appeared in the time this volume was being prepared are briefly outlined below.

The minimum-time link scheduling problem for emptying a wireless network with deadlines is considered in [19], in which fundamental results of optimality characterization for arbitrary rate region have been derived. The authors demonstrate how link deadlines can be equivalently viewed as modifications of the rate region, and derive an optimization formulation that incorporates deadlines without the need of modeling the transmission sequence explicitly.

The work in [59] considers scheduled access and slotted ALOHA-like random access. Under scheduled access the nodes take turns and they get feedback on whether a transmitted packet was received successfully by the sink. A node may transmit more than once to overcome channel uncertainty. For the slotted ALOHA-like access scheme, each node attempts to transmit in every slot with a given probability. For the scheduled access and the slotted ALOHA-like schemes AoI is derived. In particular, for the slotted ALOHA-like access, an approximation on transmission probabilities that minimize the AoI is presented.

The work in [21] considers a broadcast network, where many users are interested in different pieces of information that should be delivered by a base station. The paper focuses on long-run average Age of Information and shows that an optimal scheduling algorithm is a simple stationary switch-type. That is, given the age of all other users, an optimal decision for a user is based on an age threshold. The authors leverage an infinite state Markov decision process; however considering practical cases, a sequence of finite state approximations is proposed and shown to converge. An optimal offline and an online scheduling algorithm are given; the latter does not require arrival statistics for the problem setup investigated.

In [25], the authors consider an unreliable broadcast network, where a base station

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

sends updates to a set of clients and formulate a discrete-time decision problem to find a scheduling policy that minimizes the expected weighted sum AoI of the clients in the network. Results presented show that a Greedy Policy, which transmits the packet with the highest current age, is optimal for the case of symmetric networks.

In [24] the authors outline the inefficiency of conventional approaches in maintaining fresh information updates of multiple continuous flows, and show the critical value of both age and interarrival times. They develop a new scheduler, which can operate regardless of knowledge of arrival rates, and accounts for both age and interarrival times of incoming packets. Upper and lower age performance bounds are provided. Analytical results were obtained for heavy-traffic conditions, however numerical outcomes also support the proposed scheduler even in lightly-loaded conditions.

Here we outline the work in [4] which is treated in Subsection 4.3 and the work in [5] which is treated in Subsection 6.2 to cover all existing works in scheduling to the best of our knowledge.

5.4 Summary

In this section we chose to close the discussion of the age with three areas where AoI has been utilized as a tool to deliver novel solutions. We remind once more, at this point, that the discussion depth of different works is varying in this volume, for reasons having to do even with the freshness of the work. Indeed age is a very fast evolving topic of research and pausing to account for past works is bound to yield non-equal treatment. Clearly, we chose to list some works here that have been presented, mostly in conferences, in the last few months. The following section is entirely dedicated to such works, for which we provide also a classification into different areas of application.

6 Latest Contributions

In the preceding sections we reviewed in some detail the first wave of research that followed the introduction of the concept of age which have appeared between 2011 and 2017. The explosive growth of work on this concept that ensued since then has led to approximately 50 papers, published between 2011 and today, that have explored an amazing diversity of topics in which age plays a role. In this section, we offer a brief review if not a mere outline, of most of this work.

6.1 Effect of Errors

One common assumption in most works until 2016 is that the status update delivery is always successful. We list here a set of recent papers that take into account the case of

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

imperfect packet transmission. The first work addresses transmission policies considering PAoI in an M/M/1 queueing system with packet delivery errors. The second work leverages transmission of coded updates over the binary erasure channel. The third work examines preempting or dropping an update in a buffer-less system where transmission takes place over an erasure channel. The fourth work considers schemes utilizing the correlation in source update messages to transmit differential information to the destination. The fifth work addresses a data protection scheme examining random linear coding versus timely delivery for updates over erasure channels.

- In [8], the authors consider the Peak Age of Information (PAoI) in an M/M/1 queueing system with packet delivery errors. They consider two policies: (i) the last-come-first-served (LCFS) scheduling and (ii) to keep transmitting the most recent packet upon reception by retransmissions. Exact expressions for PAoI for both policies are derived. Both policies eliminate the queueing delay and they can ensure a small PAoI.

- The work in [60] investigates the transmission of coded updates through a binary erasure channel to a destination by using AoI. More specifically, the average age is derived for an infinite incremental redundancy (IIR) system in which the transmission of a k-symbol update continues until k symbols are received. Then, the authors compare the IIR system with a fixed redundancy (FR) system. In the FR system, an update is transmitted as a packet with n symbols and the packet is successfully received if and only if at least k symbols are received, otherwise it is discarded. For the case of single destination, the authors show that the FR system can perform as well as the IIR system, however, as the number of destinations increases the FR system outperforms the IIR system.

- In [41], a system with randomly generated updates to be transmitted to a destination is considered. However, only a single update can be in the transmission service at a time. Thus, the source needs to prioritize between two transmission policies, the preemption of the current update or discarding the new one. The authors consider Poisson arrivals and general service time, which is referred as an M/G/1/1 queue. They study the average age and the optimal update arrival rate for these two schemes under general service time distribution. Then, they apply the derived results on two scenarios where the updates are transmitted through an erasure channel. Two cases are considered for the transmission, (a) an IIR Hybrid ARQ (HARQ) system and (b) a FR HARQ system. It is shown that in both schemes the best strategy is to not preempt. In addition, it is also proven that IIR is better than FR regarding AoI.

- In [7], the authors consider data transmission schemes for a single source, sending periodic updates to a receiver through an unreliable channel. Two schemes are

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

considered that exploit the correlation in the source messages in order to transmit differential information to the receiver. In the first scheme, the source by utilizing the receiver feedback, can decide to send among the differential and the actual information at each transmission opportunity. On the other hand, in the second scheme without feedback, the source periodically transmits the actual information, interspersed with differential messages. It is observed that the differential encoding improves the timeliness performance, when there is feedback at the receiver.

- The authors in [46], consider the tradeoff between data protection and timely delivery of collected data in the context of real-time status updates over erasure channels. The authors consider random linear coding for the transmission and they show the effect of coding parameters on the performance.

6.2 Beyond Single-hop Communication

Here we list six recent works published in 2017 that address multiple communication links between the source and destination.

The first paper addresses minimization of AoI in general multihop networks. The second considers two fixed-delay hops with energy harvesting. In 5.3 we have included for completeness of the scheduling treatment there four more works. Broadcasting and scheduling over unreliable channels is addressed in [25]. The wireless broadcast network case is explored in [21]. A scheduler that addresses multiple continuous flows accounting AoI and interrarival times is presented in [24]. Finally, towards emptying a wireless network with deadlines we outline [19].

- The first attempt to address multi-hopping for status update transmission is made in [5], where the authors minimize the AoI in general multi-hop networks. The key outcome of this work is that there are simple policies which can achieve optimality of the age processes under arbitrary network topologies. If update packets do not enter the network in order, the last-generated first-served (LGFS) policy achieves smaller age, assuming exponentially distributed packet transmission times. Furthermore, for arbitrary distributions of packet transmission times, the non-preemptive LGFS policy minimizes the age processes at all nodes in the network among all non-preemptive work-conserving policies. Age-optimality here can be achieved even if the transmission time distribution differs from one link to another, i.e., the transmission time distributions are heterogeneous.

- In [1], the authors propose age optimal policies for energy harvesting two-hop networks assuming fixed transmission delays. They show that the optimal policy is that the relay's data buffer should not contain any packets waiting for service and the source

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

should send an update to the relay when it is ready to forward. This allows for treatment of the source and relay nodes as one combined node communicating with the destination node, and reduces the two-hop problem to a single hop one. Thus, the single hop problem was solved by balancing the inter-update times to the extent allowed by energy arrival times and transmission delays.

6.3 Other Topics

Aside from the two large classes we already categorized the recent works, there are more works published of late and we list them here, reminding the reader that we cover works up to mid-2017. The first two works explore fundamental relations between AoI and the related metrics of PAoI and VoUI. The third work changes the update stochastic flow from the source to the destination to a demand-based "pull" scheme. The fourth work addresses streaming source coding for optimal blocklength. The fifth work address QoE in cloud gaming optimizing for age of video frame updates. The sixth work addresses global CSI estimation and dissemination, for fully connected wireless networks with reciprocal channels. With respect to topics related to energy harvesting we refer the reader back to Subsection 5.2, where we have already outlined two works ([3] and [55]) that recently appeared.

- In [23] the stationary distribution of the AoI, is studied and an invariant relation among the distributions of the AoI, the Peak AoI, and the system delay is derived. This relation holds for a wide class of information update systems, including both FCFS and LCFS disciplines. Furthermore, for the stationary ergodic FCFS GI/GI/1 queue, the paper gives the stationary distributions of the AoI and the peak AoI in terms of the system delay distribution. Finally, explicit formulas for the Laplace-Stieltjes transforms (LSTs) of the stationary distributions of the AoI and the peak AoI, as well as the first two moments of AoI, in the stationary FCFS M/GI/1 and GI/M/1 queues are given.

- The work in [36] aimed to expand the concept of information ageing by introducing the Cost of Update Delay (CoUD) metric to provide a flexible measure of having stale information at the system monitor. The work further introduced the metric of Value of Information of Update (VoIU) capturing the reduction of CoUD upon reception of an update. Small CoUD corresponds to timely information while VoIU represents the impact of the received information in reducing the CoUD. Therefore, in a communication system it would be highly desirable to minimize the average CoUD, and at the same time maximize the average VoIU. To this end we obtain the average VoIU for the M/M/1 queue and discuss how the optimal server utilization

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

with respect to VoIU can be used combined with the CoUD average analysis. A key in the flexibility of these notions is the potential for usage of non-linear functions to represent them, giving ground to establish differentiated service classes in monitoring systems.

- The work in [48] introduced an updates "pulling model", where the destination requests the source for updates through a request replication scheme leveraging multiple servers. The authors consider a Poisson updating process and also, they assume exponentially distributed response time. A closed-form expression of the expected AoI at the user's side is derived and the optimal solution is obtained. In the AoI minimization problem under the pull model, replication schemes exhibit a unique property and capture a novel tradeoff between different levels of information freshness and different response times across the servers. The key outcome is that waiting for more than one response can be exploited to minimize the expected AoI at the user's side.

- The authors in [64], consider backlog-adaptive streaming source coding systems, the status age analysis is applied and it is assumed that the state of the channel interface is known at the encoder. It is shown that the average age can be reduced by allowing the encoder to dynamically adjust the blocklength. In addition, a scheme that enables the encoder to constrain the maximum allowed blocklength is introduced, this scheme improves the age performance. This work extends the authors' previous work in [63] where age analysis was applied to a streaming coding system using fixed-to-variable lossless coding schemes.

- The work in [62] introduces a model for cloud gaming systems[5] in order to optimize the timeliness of video frames based on an AoI metric. The authors provide a quantified representation of the user-perceived quality of experience of latency-sensitive real-time cloud-assisted gaming. In cloud gaming systems, is important to ensure "timely" updates of the players thus, AoI was employed to characterize the system performance.

- The work in [34][6], develops a general framework to quantify the staleness of global CSI in wireless networks. This framework considers the number of channel states disseminated in each packet and the additional data and overhead in each packet which contribute to the staleness of the CSI. In addition, CSI dissemination protocols are developed and their performance is evaluated by quantifying the maximum and average

[5] Cloud gaming, allows direct, on-demand video streaming of a game onto computers, consoles and mobile devices. The game itself is stored, executed, and rendered on a remote server and the user action results are streamed directly to their device(s) over the network.

[6] Notice that this work was already treated in Subsection 5.1, but is presented here in brevity, for completeness of this section.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

staleness as a function of the number of nodes in the network and the composition of each packet.

6.4 Summary

As these lines are written there is intense activity in the community on the many questions that have arisen in the cause of the last six years concerning AoI. Literally, one would have to generate "updates" of this volume very few months to capture the ongoing research and the accompanying innovations. Instead, we close this section by pointing out that in the next, and final, section of this volume we present our thoughts on some of the fundamental implications of the concept of age as a guide to possible future investigations. Should the reader need a timely updated repository of publications on and about age, Prof. Y. Sun, maintains, to our knowledge, the most well kept at his professional website: http://www.auburn.edu/~yzs0078/AoI.html.

7 Future Directions and Open Questions

There is an immense wealth of possible problems associated with the notion of timeliness as captured by age, in the form of either a performance index, a tool, or even a concept.

The importance of age as a performance index is primarily based on the desire to predict accurately. So, the notions of Effective Age, Value of Information of Update, etc. are related to minimizing age. In turn, this observation clearly connects to fundamentals aspects of information of the signal. Thus it leads naturally to connections with information theoretic questions, sampling theory, compression, and causal signal reconstruction.

In addition, age as a tool can play a role in formulations of caching problems, the IoT, social networks, and in several other applications. Finally, age may have a critical role in dynamic control systems, specifically, in the context of networks of autonomous vehicles and control (when multiple sources are involved). Similarly, in machine learning and distributed inference, which is the theoretical underpinning of autonomous vehicle control, the relative values of the age of signals that are received and that lead to enabling actuators to implement actions are of crucial importance. That is, whether the age of source 1 is less than or greater than that of source 2 (observing a different sensor) can make a huge difference as far as the consequences of control strategies.

7.1 Connection with Information Theory

Connection of Age of Information of a source with entropy and mutual information will be of importance and also it can lead to interesting studies regarding the predictability of a

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

source. In addition, differential transmission schemes based on feedback or knowledge of the source or the destination can decrease the amount of transmitted information inside a network without sacrificing the accuracy of the knowledge. This is becoming quite important since networks are rapidly growing in size and congestion levels are also increasing.

The need to deliver status updates, messages in general, in a timely manner prohibits the use of long codewords. Thus, the application of coding strategies with low latency is of major importance. In addition, when messages are transmitted through unreliable wireless channels, longer codewords can offer smaller probabilities of failure but longer latency. The derivation of optimal codeword length for minimizing the average AoI is an important research direction that relates to recent advances in information and coding theory.

In addition to the connections of AoI with information theory, age is a perfect example of introducing **context** in Information, Communication, and Control Systems. As mentioned in the Shannon lecture at the IEEE ISIT 2017, by D. Tse, part of which can also be found in the IEEE IT Newsletter September 2017 [53], the *context* in message transmission has so far been overlooked. In [50] is was stated that

"The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently messages have meaning... These semantic aspects are irrelevant to the engineering problem."

In addition to setting the strict limits on which the development of information theory was based, Shannon's quote above eliminated an important part of the communication process. Although "meaning" and "content" of information still elude a formal incorporation into the theoretical investigation of the communication process, "context" is not. That is, communication may occur in order to *control* or to *compute* or to *infer* and not just to reproduce messages chosen at the source. Age of Information is a perfect example of context introduction. In fact, depending on the context age can be modified to migrate to an Effective Age, different for each application. This we view as an important future challenge.

7.2 Connection with Signal Processing

As discussed earlier, making decisions in networked control and monitoring systems is based on the timeliness of the delivered information. Thus, a fundamental part of sampling of the source signal becomes its timeliness requirement. Unavoidably, this draws a clear connection between the concept of AoI and the theory of sampling for signal reconstruction. Moreover, in the same context, it is important to reconstruct correctly source messages from received samples, that have been delivered through an erroneous/noisy channel. This problem and its connection with AoI is a first direction to be considered that can lead to significant results.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

7.2.1 Sampling and Remote Estimation

From the first publication in this direction [51], it was proven that for a Wiener process sampling that minimizes AoI is not optimal for prediction. Rather, the optimal sampling policy is a threshold-based one. However, as proven in that work, if the sampling times are independent of the observed Wiener process, the optimal sampling problem reduces to an AoI optimization problem. A step further is to consider more realistic signal models, such as, for example, the Ornstein–Uhlenbeck process [54]. The Ornstein–Uhlenbeck process can be considered as the continuous-time analogue of the discrete-time AR(1) process.

In addition, it is important to consider the case where the samples are transmitted through multiple paths and arrive with random delays at the remote receiver. The latter describes more realistic scenarios in wireless transmissions. The problem of optimal sampling and remote reconstruction can be also considered with the by-product metrics of AoI such as the VoIU that we recently defined. Clearly, there are a lot of research opportunities towards this direction. For example, considering the delay/deadline constrained remote reconstruction of a source is a problem that has not been considered yet in the current literature, extending the work in [28]. This can be motivated by scenarios where the information needed to be received and reconstructed at a remote receiver has specific delay or deadline constraints, such as real-time critical information, for example a remote surgery.

If the process under observation consists of highly correlated data, then the frequency of transmission of updates can be significantly reduced without affecting the timeliness of the information at a remote receiver. Consider for example the case of a constant process, then the transmission of a single sample is sufficient to sustain zero AoI. Thus, defining the Effective Age and connecting it with the autocorrelation properties of the source under observation is of major importance. In addition, the relation of Effective Age with AoI and its by-product metrics call for further investigations.

7.2.2 Sampling and Information Theory

A fundamental part of the AoI concept is the process of sampling of the source signal. Unavoidably, this draws a clear connection to the theory of sampling for signal reconstruction. Nyquist sampling theory needs to be extended to apply to causal reconstruction. Namely, most of the emphasis in sampling-related research in signal processing considers signal reconstruction that is not limited to causal samples. An important, unexplored issue, is whether, and to what degree, can we reconstruct signals based only on past and/or delayed samples. There is an obvious connection between sampling and information theory that has been explored by Y. Eldar and A. Goldsmith, among others, in [9], [15]. Going beyond that, the question arises whether the age of information, after reception of a signal that introduced random delay plays a role in the reconstructability of the signal. This question is different than the one that merely considers prediction quality as it involves deeper issues

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

of signal structure and sampling patterns.

7.3 Applications

In this part, we address a couple of more practical directions where AoI and the by-product metrics can be utilized. We clearly do not cover all possible fields of application for AoI as a tool, giving only the broad areas of content caching and machine learning based data analytics.

7.3.1 Caching

An immediately promising area for work lies in caching and content placement in general. Already there have been the first works on caching policies aiming to minimizing cache misses [29] using the request rates and popularity, based on content age, and updating dynamic content in a cache in order to minimize the average age of cached items [57]. Both these works deliver optimal policies for their models, however, it is directly indicated in them that more general cases (generic request models that can for example model viral content) are not tractable, even for toy-problem models. Caching of messages has lately become of critical importance due to the pervasiveness of highly demanding online services (such as online gaming, video on demand, and augmented reality applications) and the IoT.

7.3.2 Data Analytics

Data analytics, based on machine learning (ML), relies on the so called data analytics pipeline [37], which is a string of data and process manipulations aimed at making ML cope with Big Data [38]. One of the crucial steps in that pipeline is the so called "processing manipulations" which focuses on modifying how data are processed and stored. Processing techniques can take advantage of the inherent parallel nature of algorithms to provide significant performance improvement. Within those, the horizontal scaling paradigm refers to distributed systems where processing is dispersed over networked nodes. Involving distributed nodes requires network communication and coordination. Such horizontal scaling systems can be categorized as Batch- and Stream- oriented ones. The latter operate on small sets of recent data in real-time or near real-time. Developers of algorithmic trading, fraud detection, and surveillance systems have been especially interested in such solutions [16]. The importance of real-time processing of data from sensors, mobile devices, and IoT deployments has resulted in the emergence of a number of streaming systems; examples include Twitter's Storm [39] and Yahoo's S4 [43]. The importance of the Age of Information here is evident as the volume of data input can significantly be decreased, should certain data sources provide non-timely data.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

7.4 Summary

As we have seen in this section, AoI and the notions of Peak Age, Value of Information of Update, and so forth can be tools with which one can open up new areas in well-established domains. One can also use these as metrics and tools and address problems with a new set of requirements and defining constraints that are themselves timely in view of the evolving trends in communication, where the context is shifting from the mere transport of bits to serving tasks such as estimation, reconstruction, data dissemination and storage to name a few.

Acknowledgements

We would like to thank Prof. Tony Ephremides for multiple discussions, insightful comments, and most helpful steering, in the preparation of this volume.

The authors would like to acknowledge the support of the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 642743 (WiVi-2020).

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

References

[1] Arafa, A. and S. Ulukus. 2017. "Age-Minimal Transmission in Energy Harvesting Two-hop Networks". *CoRR*. abs/1704.08679. URL: http://arxiv.org/abs/1704.08679.

[2] Bacinoglu, B. T., E. T. Ceran, and E. Uysal-Biyikoglu. 2015. "Age of information under energy replenishment constraints". In: *Information Theory and Applications Workshop (ITA)*. 25–31.

[3] Bacinoglu, B. T. and E. Uysal-Biyikoglu. 2017. "Scheduling status updates to minimize age of information with an energy harvesting sensor". In: *IEEE International Symposium on Information Theory (ISIT)*. 1122–1126.

[4] Bedewy, A. M., Y. Sun, and N. B. Shroff. 2016. "Optimizing data freshness, throughput, and delay in multi-server information-update systems". In: *IEEE International Symposium on Information Theory (ISIT)*. 2569–2573.

[5] Bedewy, A. M., Y. Sun, and N. B. Shroff. 2017. "Age-optimal information updates in multihop networks". In: *IEEE International Symposium on Information Theory (ISIT)*. 576–580.

[6] Bertsekas, D. and R. Gallager. 1992. *Data Networks (2Nd Ed.)* Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

[7] Bhambay, S., S. Poojary, and P. Parag. 2017. "Differential Encoding for Real-Time Status Updates". In: *IEEE Wireless Communications and Networking Conference (WCNC)*. 1–6.

[8] Chen, K. and L. Huang. 2016. "Age-of-information in the presence of error". In: *IEEE International Symposium on Information Theory (ISIT)*. 2579–2583.

[9] Chen, Y., Y. C. Eldar, and A. J. Goldsmith. 2013. "Shannon Meets Nyquist: Capacity of Sampled Gaussian Channels". *IEEE Transactions on Information Theory*. 59(8): 4889–4914.

[10] Cooper, R. B. 1981. *Introduction to queueing theory*. Includes index. New York: North Holland.

[11] Costa, M., M. Codreanu, and A. Ephremides. 2014. "Age of information with packet management". In: *IEEE International Symposium on Information Theory (ISIT)*. 1583–1587.

[12] Costa, M., M. Codreanu, and A. Ephremides. 2016. "On the Age of Information in Status Update Systems With Packet Management". *IEEE Transactions on Information Theory*. 62(4): 1897–1910.

[13] Costa, M., S. Valentin, and A. Ephremides. 2015a. "On the age of channel information for a Finite-State Markov model". In: *IEEE International Conference on Communications (ICC)*. 4101–4106.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

[14] Costa, M., S. Valentin, and A. Ephremides. 2015b. "On the age of Channel State Information for non-reciprocal wireless links". In: *IEEE International Symposium on Information Theory (ISIT)*. 2356–2360.

[15] Eldar, Y. C. 2015. *Sampling Theory: Beyond Bandlimited Systems*. Cambridge University Press.

[16] Grolinger, K., M. Hayes, W. A. Higashino, A. L'Heureux, D. S. Allison, and M. A. M. Capretz. 2014. "Challenges for MapReduce in Big Data". In: *IEEE World Congress on Services*. 182–189.

[17] He, Q., D. Yuan, and A. Ephremides. 2016a. "On optimal link scheduling with min-max peak age of information in wireless systems". In: *IEEE International Conference on Communications (ICC)*. 1–7.

[18] He, Q., D. Yuan, and A. Ephremides. 2016b. "Optimizing freshness of information: On minimum age link scheduling in wireless systems". In: *14th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*. 1–8.

[19] He, Q., D. Yuan, and A. Ephremides. 2017. "On optimal link scheduling with deadlines for emptying a wireless network". In: *IEEE International Symposium on Information Theory (ISIT)*. 461–465.

[20] Hespanha, J. P. 2006. "Modelling and analysis of stochastic hybrid systems". *IEE Proceedings - Control Theory and Applications*. 153(5): 520–535.

[21] Hsu, Y. P., E. Modiano, and L. Duan. 2017. "Age of information: Design and analysis of optimal scheduling algorithms". In: *IEEE International Symposium on Information Theory (ISIT)*. 561–565.

[22] Huang, L. and E. Modiano. 2015. "Optimizing age-of-information in a multi-class queueing system". In: *IEEE International Symposium on Information Theory (ISIT)*. 1681–1685.

[23] Inoue, Y., H. Masuyama, T. Takine, and T. Tanaka. 2017. "The stationary distribution of the age of information in FCFS single-server queues". In: *IEEE International Symposium on Information Theory (ISIT)*. 571–575.

[24] Joo, C. and A. Eryilmaz. 2017. "Wireless scheduling for information freshness and synchrony: Drift-based design and heavy-traffic analysis". In: *15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*. 1–8.

[25] Kadota, I., E. Uysal-Biyikoglu, R. Singh, and E. Modiano. 2016. "Minimizing the Age of Information in broadcast wireless networks". In: *54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. 844–851.

[26] Kam, C., S. Kompella, and A. Ephremides. 2014. "Effect of message transmission diversity on status age". In: *IEEE International Symposium on Information Theory (ISIT)*. 2411–2415.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

[27] Kam, C., S. Kompella, G. D. Nguyen, and A. Ephremides. 2016a. "Effect of Message Transmission Path Diversity on Status Age". *IEEE Transactions on Information Theory*. 62(3): 1360–1374.

[28] Kam, C., S. Kompella, G. D. Nguyen, J. E. Wieselthier, and A. Ephremides. 2016b. "Age of information with a packet deadline". In: *IEEE International Symposium on Information Theory (ISIT)*. 2564–2568.

[29] Kam, C., S. Kompella, G. D. Nguyen, J. E. Wieselthier, and A. Ephremides. 2017. "Information freshness and popularity in mobile caching". In: *IEEE International Symposium on Information Theory (ISIT)*. 136–140.

[30] Kaul, S. K., R. D. Yates, and M. Gruteser. 2012a. "Status updates through queues". In: *46th Annual Conference on Information Sciences and Systems (CISS)*. 1–6.

[31] Kaul, S., M. Gruteser, V. Rai, and J. Kenney. 2011a. "Minimizing age of information in vehicular networks". In: *8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*. 350–358.

[32] Kaul, S., R. Yates, and M. Gruteser. 2011b. "On Piggybacking in Vehicular Networks". In: *IEEE Global Telecommunications Conference (GLOBECOM)*. 1–5.

[33] Kaul, S., R. Yates, and M. Gruteser. 2012b. "Real-time status: How often should one update?" In: *IEEE International Conference on Computer Communications (INFOCOM)*. 2731–2735.

[34] Klein, A. G., S. Farazi, W. He, and D. R. Brown. 2017. "Staleness Bounds and Efficient Protocols for Dissemination of Global Channel State Information". *IEEE Transactions on Wireless Communications*. 16(9): 5732–5746. ISSN: 1536-1276.

[35] Kleinrock, L. 1975. *Queueing Systems*. Vol. I: Theory. Wiley Interscience.

[36] Kosta, A., N. Pappas, A. Ephremides, and V. Angelakis. 2017. "Age and value of information: Non-linear age case". In: *IEEE International Symposium on Information Theory (ISIT)*. 326–330.

[37] Labrinidis, A. and H. V. Jagadish. 2012. "Challenges and opportunities with big data". *Proceedings of the VLDB Endowment*. 5(12): 2032–2033.

[38] L'Heureux, A., K. Grolinger, H. F. Elyamany, and M. A. M. Capretz. 2017. "Machine Learning With Big Data: Challenges and Approaches". *IEEE Access*. 5: 7776–7797.

[39] Marz, N. 2014. *Apache Strom*. URL: http://storm.apache.org.

[40] Najm, E. and R. Nasser. 2016. "Age of information: The gamma awakening". In: *IEEE International Symposium on Information Theory (ISIT)*. 2574–2578.

[41] Najm, E., R. Yates, and E. Soljanin. 2017. "Status updates through M/G/1/1 queues with HARQ". In: *IEEE International Symposium on Information Theory (ISIT)*. 131–135.

[42] Nelson, R. 1995. *Probability, stochastic processes and queueing theory: the mathematics of computer performance modeling*. New York: Springer-Verlang.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

[43] Neumeyer, L., B. Robbins, A. Nair, and A. Kesari. 2010. "S4: Distributed Stream Computing Platform". In: *IEEE International Conference on Data Mining Workshops*. 170–177.

[44] Papoulis, A. 1991. *Probability, random variables, and stochastic processes. McGraw-Hill series in electrical engineering*. New York: McGraw-Hill.

[45] Pappas, N., J. Gunnarsson, L. Kratz, M. Kountouris, and V. Angelakis. 2015. "Age of information of multiple sources with queue management". In: *IEEE International Conference on Communications (ICC)*. 5935–5940.

[46] Parag, P., A. Taghavi, and J. F. Chamberland. 2017. "On Real-Time Status Updates over Symbol Erasure Channels". In: *IEEE Wireless Communications and Networking Conference (WCNC)*. 1–6.

[47] Rudin, W. 1987. *Real and Complex Analysis, 3rd Ed.* New York, NY, USA: McGraw-Hill, Inc. ISBN: 0070542341.

[48] Sang, Y., B. Li, and B. Ji. 2017. "The Power of Waiting for More than One Response in Minimizing the Age-of-Information". *CoRR*. abs/1704.04848. URL: http://arxiv.org/abs/1704.04848.

[49] Sesia, S., I. Toufik, and M. Baker. 2011. *LTE, The UMTS Long Term Evolution: From Theory to Practice*. Wiley Publishing.

[50] Shannon, C. 1948. "A mathematical theory of communication". *Bell system technical journal*. 27.

[51] Sun, Y., Y. Polyanskiy, and E. Uysal-Biyikoglu. 2017. "Remote estimation of the Wiener process over a channel with random delay". In: *IEEE International Symposium on Information Theory (ISIT)*. 321–325.

[52] Sun, Y., E. Uysal-Biyikoglu, R. Yates, C. E. Koksal, and N. B. Shroff. 2016. "Update or wait: How to keep your data fresh". In: *35th Annual IEEE International Conference on Computer Communications (INFOCOM)*. 1–9.

[53] Tse, D. 2017. "2017 Shannon Lecture: The Spirit of Information Theory". In: *IEEE Information Theory Society Newsletter*.

[54] Uhlenbeck, G. E. and L. S. Ornstein. 1930. "On the theory of Brownian Motion". *Phys. Rev.* 36: 823–841.

[55] Wu, X., J. Yang, and J. Wu. 2017. "Optimal status updating to minimize age of information with an energy harvesting source". In: *IEEE International Conference on Communications (ICC)*. 1–6.

[56] Yates, R. D. 2015. "Lazy is timely: Status updates by an energy harvesting source". In: *IEEE International Symposium on Information Theory (ISIT)*. 3008–3012.

[57] Yates, R. D., P. Ciblat, A. Yener, and M. Wigger. 2017a. "Age-optimal constrained cache updating". In: *IEEE International Symposium on Information Theory (ISIT)*. 141–145.

The final version of record of this article is published in:
Antzela Kosta, Nikolaos Pappas and Vangelis Angelakis (2017),
"Age of Information: A New Concept, Metric, and Tool",
Foundations and Trends® in Networking: Vol. 12: No. 3, pp 162-259.
http://dx.doi.org/10.1561/1300000060

[58] Yates, R. D. and S. Kaul. 2012. "Real-time status updating: Multiple sources". In: *IEEE International Symposium on Information Theory (ISIT)*. 2666–2670.

[59] Yates, R. D. and S. K. Kaul. 2017. "Status updates over unreliable multiaccess channels". In: *IEEE International Symposium on Information Theory (ISIT)*. 331–335.

[60] Yates, R. D., E. Najm, E. Soljanin, and J. Zhong. 2017b. "Timely updates over an erasure channel". In: *IEEE International Symposium on Information Theory (ISIT)*. 316–320.

[61] Yates, R. D. and S. K. Kaul. 2016. "The Age of Information: Real-Time Status Updating by Multiple Sources". *CoRR*. abs/1608.08622. URL: http://arxiv.org/abs/1608.08622.

[62] Yates, R. D., M. Tavan, Y. Hu, and D. Raychaudhuri. 2017c. "Timely Cloud Gaming". In: *IEEE International Conference on Computer Communications (INFOCOM)*. 2286–2294.

[63] Zhong, J. and R. D. Yates. 2016. "Timeliness in Lossless Block Coding". In: *2016 Data Compression Conference (DCC)*. 339–348.

[64] Zhong, J., R. D. Yates, and E. Soljanin. 2017. "Backlog-adaptive compression: Age of information". In: *IEEE International Symposium on Information Theory (ISIT)*. 566–570.