Statistical Delay Bounds for Automatic Repeat Request Protocols with Pipelining

Mark Akselrod and Markus Fidler
Institute of Communications Technology
Leibniz Universität Hannover

***Abstract*—The recent trend towards low-latency wireless communication requires a notion of non-ergodic capacity that deals with delays. Significant research in areas, such as effective capacity, delay-constrained capacity, and stochastic service curves, has contributed such results for relevant physical layer aspects, like fading processes, interference, and multiple antenna systems. Less attention has been paid to actual implementations of link layer automatic repeat request protocols. Instead, error-free transmission using instantaneous channel state information, simple stop-and-wait protocols, or instantaneous feedback are frequently assumed. In this work, we investigate protocols with pipelining that deal with packet errors under non-negligible round-trip-times. We define a stochastic service curve model of a general class of automatic repeat request protocols with pipelining and derive statistical waiting time and sojourn time bounds. We discover two regimes: under low to moderate load retransmissions cause the largest part of the sojourn time, whereas under high load the waiting time dominates the sojourn time. Generally, with increasing round-trip-time the basic cases of stop-and-wait protocols or instantaneous feedback neglect relevant effects and provide less accurate estimates.**

I. Introduction

Today's cellular packet data networks employ advanced link layer automatic repeat request (ARQ) protocols to ensure high throughput even in the case of frequent transmission errors. The throughput is optimized using a large degree of pipelining to avoid stalling, combined with positive and negative acknowledgements for selective repeat of erroneous packets. In addition, cellular networks control the packet error rate by selection of the modulation order and the coding rate of forward error correction from a range of modulation and coding schemes. These enable trading payload bits for a more robust transmission. Given this tradeoff, the packet error rate can be adjusted to maximize the long-term average throughput.

Beyond the long-term throughput, that has been of primary concern, the recent trend towards low-latency wireless communications shifts the focus towards delays. The stochastic nature or wireless channels entails capacity fluctuations that are considered by non-ergodic capacity models, such as outage capacity [1], delay-limited capacity [2], and effective capacity [3], [4]. While capacity fluctuations can cause queueing delays at the sender, additional protocol delays arise in case of packet loss. These protocol delays, that are due to feedback, timeouts, retransmissions, as well as reordering of out of sequence packets at the receiver, have been investigated to a much lesser extent.

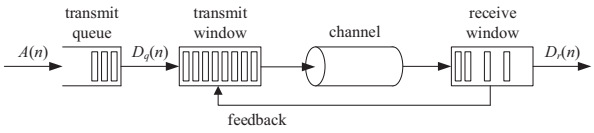

Fig. 1. ARQ protocol with pipelining and selective repeat.

A common approach is to express the time-varying capacity of a wireless channel by a random service process. Alternatively, [5], [6] consider the capacity of a channel under ideal conditions and define a random impairment process that is deducted afterwards to derive a residual service process. This equivalent view makes a connection with models known in scheduling. Recent works have contributed characterizations of the service processes of, e.g., fading channels [7]–[9], finite block-length codes [10], [11], hybrid-ARQ [12], interference channels [13], cognitive radio [14], [15], multiple-input multiple-output (MIMO) [16], and multi-access channels [17].

An important difference in wireless channel models is the type of channel state information that is used. In case of instantaneous channel state information, the sender is assumed to adapt the transmission rate to the instantaneous channel capacity so that it can avoid any transmission errors. In this case, the service process is determined as the instantaneous channel capacity [9], [13]. On the other hand, if only statistical channel state information is given, the sender can tune the transmission rate, e.g., to maximize the average throughput, but it can generally not avoid transmission errors. The result is an on-off service process where the transmission is either successful or not, i.e., the channel is in on or in off state, respectively. Channels of the on-off type and variations thereof are investigated in [7], [8], [10]–[12], [14]–[17].

Another common assumption is that packets are generally transmitted in sequence, i.e., packet $n+1$ is transmitted only after packet n is successfully delivered, respectively, acknowledged. In the case of the on-off channel model, this assumption applies likewise in case of stop-and-wait ARQ, if instantaneous feedback about the success of transmissions is available, or if the sender has instantaneous channel state information so that it can pause the transmission whenever the channel is in off state. It simplifies the analysis considerably, since it avoids any difficulties that are due to feedback delays and reordering of packets in more advanced ARQ protocols. Basically all of the above works fall into this category.

978-0-9883045-5-0/18/$31.00 ©2018 ITC
DOI 10.1109/ITC30.2018.00011

ARQ protocols with pipelining and non-negligible feedback delays are considered in [18]. The authors model the feedback control loop of ARQ by a system of self-dependent equations that define retransmission flows $i \in 1, \ldots, N-1$ and derive a fixed-point solution. Retransmission flow i comprises only those packets that are retransmitted for the ith time and N is the maximum number of transmission attempts. The retransmission flows are superposed assuming that retransmission flows have increasing priority with i. This implies that older packets are retransmitted first. It does, however, not ensure that data are delivered in sequence. Delays that are due to resequencing of data are not considered.

In this work, we define a model of a general ARQ protocol with pipelining and selective repeat as depicted in Fig. 1. Compared to [18], that uses a space-domain model, where cumulative data arrivals are expressed as functions of time, we use a time-domain approach, that uses sequences of packet timestamps indexed by the packet number. This enables expressing the reordering of out of sequence packets by a maximum operation of packet timestamps, which is not easily possible in the space-domain model. For an introduction to space-domain and time-domain models see the textbooks [3], [19], and [20] that shows the isomorphism of the underlying algebras of the two domains. Using the time-domain branch of the stochastic network calculus, as in e.g., [21]–[23], we derive stochastic bounds of the waiting time and the sojourn time of the ARQ protocol. Our numerical evaluation reveals two regimes: under low to moderate load, the tail distribution of the waiting time decays quickly and the sojourn time is dominated by retransmission delays that are not considered by the related works; under high load, on the other hand, the waiting time prevails. We also include simulation results to verify the accuracy of the bounds.

The remainder of this paper is structured as follows. In Sec. II, we define the ARQ protocol and model its service process. Stochastic delay bounds for general traffic arrivals are derived in Sec. III. In Sec. IV we discuss numerical results, and in Sec. V, we derive delay bounds for arrivals with independent increments and show the tightness of the bounds compared to simulation results. Brief conclusions are provided in Sec. VI.

II. ARQ Protocol Model

In this section, we first define the ARQ protocol before we deduce the corresponding service process. The service process is used to derive the packet departure timestamps of the ARQ protocol from its packet arrival timestamps.

A. Protocol Definition

We consider an automatic repeat request protocol with pipelining and selective repeat. Packets are transmitted one after another in first-come first-served (fcfs) order. The transmission of new packets is suspended to schedule retransmissions of earlier packets as needed. If the transmission of a packet is successful, the receiver selectively acknowledges that packet. The acknowledgement is available at the sender after one round-trip-time t_R. Otherwise if there is no acknowledgement, e.g., due to bit errors or packet loss, a retransmission is triggered by a timeout, $t_O \geq t_R$. This ensures that at any time there is at most one copy of any packet in transmission or awaiting the acknowledgement, respectively.

To implement the protocol, the sender maintains two buffers: a fcfs transmit queue where packets that await their first transmission are queued, and a transmit window where the packets that have been transmitted and await acknowledgement are stored for potential retransmission. A packet that is acknowledged is removed from the transmit window. The receiver maintains a receive window where it buffers packets that are received out of sequence. The receiver waits until gaps in the sequence are filled by retransmissions of missing packets before it delivers packets in their original sequence.

We assume a slotted transmission system that transmits packets of a fixed length l. Given capacity c, the slot time is determined as the transmission time of a packet $t_T = l/c$. Time-slots are numbered by $i \geq 1$ where the interval $((i-1)t_T, it_T]$ is the ith time-slot. The propagation delay from the sender to the receiver is given as $t_P = d/s$ where d is the distance and s the speed of signal propagation.

Packet transmissions are successful with probability p_P and are statistically independent of each other. Independence is achieved by transmission systems for example by use of frequency hopping. Similarly, an acknowledgement is successfully delivered with probability p_K. Hence, t_R units of time after transmitting a packet, the sender receives an acknowledgement with probability $p_R = p_P p_K$.

Since packets are either acknowledged and removed from the transmit window after t_R or retransmitted after $t_O \geq t_R$, there can be at most $w = \lceil t_O/t_T \rceil$ packets in the transmit window. We choose the smallest timeout $t_O \geq t_R$ that is an integer multiple of the slot time, i.e, $t_O = \lceil t_R/t_T \rceil t_T$. Hence, if the acknowledgement fails to appear within t_R units of time, the packet is immediately retransmitted in the next time-slot due to a timeout. Consequently, if there was a packet transmission in time-slot $i \geq 1$, then in time-slot $i + w$ the packet is either acknowledged and a new packet can be transmitted or a timeout occurred for that packet and it is retransmitted. If there was no packet transmission in time-slot $i \geq 1$, then there is neither an acknowledgement nor a timeout in time-slot $i+w$ so that the time-slot is generally available for transmission of a new packet. An example is shown in Fig. 2, where packet 1 is acknowledged within t_R whereas packet 2 is lost and retransmitted after t_O and so on.

We denote the time of arrival of packet $n \geq 1$, i.e., the time when the packet is placed into the sender's transmit queue, by $A(n)$. By convention $A(0) = 0$ and $A(n) \geq A(n-1)$ for $n \geq 1$. We use shorthand notation $A(m, n) = A(n) - A(m)$ for $n \geq m \geq 1$ to denote the inter-arrival time between packet m and n. Similarly, $D(n)$ is the departure time of packet n. We distinguish the departure time from the transmit queue $D_q(n)$ and the departure time from the receive window $D_r(n)$, see Fig. 1. We assume that arrivals are processed at time-slot boundaries, i.e., $A(n) \in \{it_T, i \geq 0\}$. Otherwise given $A^c(n)$

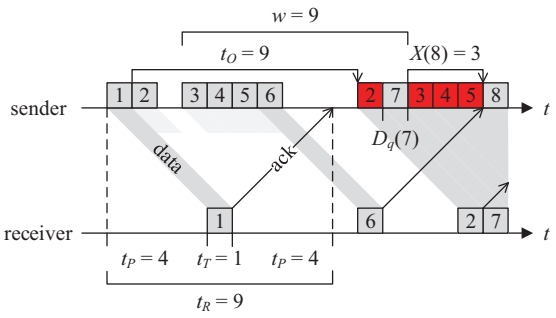

Fig. 2. Example time sequence of the ARQ protocol. $X(n+1) \geq 0$ counts the number of time-slots after $D_q(n)+1$ until the next time-slot is available for transmission of a new packet. E.g., time-slot $D_q(7)+1$ is not available for packet 8 since retransmissions of packets 3, 4, and 5 are scheduled. Hence $X(8) = 3$ and packet 8 is transmitted at $D_q(7) + 1 + X(8)$.

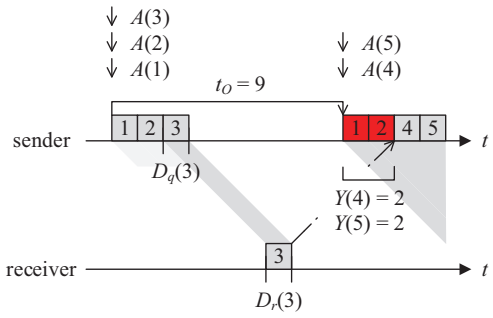

Fig. 3. Example time sequence of the ARQ protocol. $Y(n) \geq 0$ counts the number of time-slots starting at $A(n)$ until the next time-slot is available for transmission of a new packet. Although $A(4) > D_q(3) + 1 + X(4)$ packet $n = 4$ cannot be transmitted at $A(4)$ but only at $A(4) + Y(4)$ since retransmissions of earlier packets are scheduled.

in continuous time, we use $A(n) = \lceil A^c(n)/t_T \rceil t_T$.

To avoid ambiguity, we make the convention that at time-slot boundaries it_T for $i \geq 0$ the sender implements the following order of actions: 1.) arrivals are placed into the transmit queue; 2.a) in case of an acknowledgement, the respective packet is removed from the transmit window and the timer is cleared; 2.b) in case of a timeout, the respective packet is retransmitted and a timer is started; 3) if there is no retransmission, a new packet is removed from the sender's transmit queue (if any), transmitted, stored in the transmit window, and a timer is started.

For notational simplicity, we normalize $t_T = 1$ in the following. Further, we use $t_R = t_T + 2t_P$, assuming acknowledgements have a negligible transmission time.

B. Service Process

We use the concept of max-plus server [3] to model the arrival-departure relation of the ARQ protocol in time-domain.

Definition 1 (Max-plus Server)**.** *Given a system with arrivals $A(n)$ and departures $D(n)$ for $n \geq 1$. The system is a max-plus server with service process $S(m,n)$ for $n \geq m \geq 1$ if it holds for all $n \geq 1$ that*

$$D(n) \leq \max_{m \in [1,n]} \{A(m) + S(m,n)\} =: A \otimes S(n).$$

The operator $\otimes$ denotes the convolution operation under a max-plus algebra. The max-plus convolution is order preserving, i.e., if $S(m,n)$ satisfies Def. 1 any service process $S'(m,n) \geq S(m,n)$ for $n \geq m \geq 1$ satisfies Def. 1, too.

Before we derive the service process of the ARQ protocol in Lem. 1, we make some necessary definitions. Denoting $D_q(n)$ the departures from the transmit queue, packet $n \geq 1$ is transmitted for the first time in time-slot $D_q(n)$. Starting at $D_q(n)+1$, let $X(n+1) \geq 0$ count the number of time-slots until the next time-slot becomes available for transmission of a new packet, i.e., packet $n+1$. An example is shown in Fig. 2, where after transmission of packet 7 there are three retransmissions of packets 3, 4, and 5 so that $X(8) = 3$, i.e., time-slot $D_q(7) + 1 + X(8)$ is available for transmission of packet 8 in fcfs order.

Similarly, let $Y(n) \geq 0$ count the number of time-slots starting at $A(n)$ until the next time-slot becomes available for transmission of a new packet. Note that $A(n)+Y(n)$ does not necessarily mark the first transmission of packet n as there may be older packets in the transmit queue that still wait for transmission at $A(n)+Y(n)$. Fig. 3 gives an example where $A(4) = A(5)$ and $Y(4) = Y(5) = 2$ due to retransmissions of packets 1 and 2 at $A(4)$ and $A(4)+1$. Here, packet 4 is transmitted at $A(4) + Y(4) > D_q(3) + 1 + X(4)$, whereas packet 5 is transmitted at $D_q(4) + 1 + X(5) > A(5) + Y(5)$.

Further, we define an indicator variable $I(i)$ that denotes whether time-slot $i \geq 1$ is used or not, i.e., $I(i)$ is one, if time-slot i is used for transmission, and zero otherwise. We use the convention that $I(i) = 0$ for $i \leq 0$. Now consider packet n that is removed from the transmit queue, transmitted, and put into the transmit window at $D_q(n)$. At this time, the transmit window comprises a number of unacknowledged packets with index $\nu < n$, that were transmitted during the past $w-1$ time-slots, and packet n. These packet transmissions are indicated by $I(D_q(n) - j)$ for $j \in [0, w-1]$. If $I(D_q(n) - j) = 1$, there was a packet transmission in time-slot $D_q(n) - j$ and we denote $Z(n,j) \geq 0$ the number of subsequent retransmissions that are required by that packet. Otherwise if $I(D_q(n) - j) = 0$, we let $Z(n,j) = 0$.

The following lemma states the service processes of the ARQ protocol with respect to the departure process from the transmit queue $D_q(n)$ and the departure process from the receive window $D_r(n)$, respectively.

Lemma 1 (ARQ Service Process)**.** *Given the ARQ protocol with transmit window w. For $n \geq m \geq 1$ define*

$$S_q(m,n) = \min\{w-1, Y(m)\} + n - m + \sum_{\nu=m+1}^{n} X(\nu),$$

$$S_r(m,n) = S_q(m,n) + t_P + 1 + \max_{j \in [0, w-1]} \{Z(n,j) t_O - j\}.$$

It holds for $n \geq 1$ that:

i) $D_q(n) \leq A \otimes S_q(n)$,

ii) $D_r(n) \leq A \otimes S_r(n)$.

Proof. i) For $n = 1$, packet 1 arrives at an empty system at $A(1)$. Consequently, the time until the next time-slot is available for transmission of a new packet is $Y(1) = 0$ and $S_q(1,1) = 0$. Packet 1 is immediately removed from the transmit queue and transmitted, i.e., $D_q(1) = A(1)$. Hence, the first statement of the lemma is proven for $n = 1$.

For $n \geq 2$, the first transmission opportunity of packet n is $D_q(n-1) + 1 + X(n)$. If packet n arrives later, i.e., $A(n) > D_q(n-1) + 1 + X(n)$, it holds that

$$D_q(n) = A(n) + Y(n),$$

since packet n is the first packet in the transmit queue. Further, because time-slot $D_q(n-1)+1+X(n)$ is not used in this case, there cannot be a timeout in time-slot $D_q(n-1)+1+X(n)+w$ and the time-slot is generally available for transmission of a new packet. Further, if time-slot $D_q(n-1)+1+X(n)+w$ is not used, time-slot $D_q(n-1)+1+X(n)+2w$ is available and so on. It follows that $Y(n)$ is bounded by $w-1$ if $A(n) > D_q(n-1) + 1 + X(n)$ so that we can also write

$$D_q(n) = A(n) + \min\{w-1, Y(n)\}. \tag{1}$$

Otherwise, if $A(n) \leq D_q(n-1) + 1 + X(n)$, it holds that

$$D_q(n) = D_q(n-1) + 1 + X(n). \tag{2}$$

Combining Eqs. (1) and (2), we have

$$D_q(n) \leq \max\{A(n) + \min\{w-1, Y(n)\}, D_q(n-1) + 1 + X(n)\}. \tag{3}$$

By recursive insertion of Eq. (3) and using that $D_q(1) = A(1)$ since $Y(1) = 0$, it follows for $n \geq 1$ that

$$D_q(n) \leq \max_{m\in[1,n]} \Big\{ A(m) + \min\{w-1, Y(m)\} + n - m + \sum_{\nu=m+1}^{n} X(\nu) \Big\},$$

which proves that $S_q(m,n)$ is a max-plus server.

ii) To derive $D_r(n)$ note that at $D_q(n)$ there may already be up to $w-1$ unacknowledged packets with index $\nu < n$ in the transmit window that have to be delivered before packet n. The packet that was transmitted at $D_q(n) - j$ for $j \in [0, w-1]$ requires one time-slot for transmission, t_P units of time to propagate to the receiver, plus $Z(n,j) \geq 0$ retransmissions that take t_O units of time each. Since all packets $\nu < n$ have to be delivered before packet n, it follows that

$$D_r(n) = \max_{j\in[0,w-1]} \{D_q(n) - j + t_P + 1 + Z(n,j)t_O\}. \tag{4}$$

By insertion of $D_q(n) \leq A \otimes S_q(n)$ into Eq. (4) and after reordering the maxima the second part is proven. □

Next, we provide bounds of the random variables X, Y, and Z. A random variable X' is stochastically greater than a random variable X, if it holds that $\mathsf{P}[X' > x] \geq \mathsf{P}[X > x]$ for all x. In this case, we write $X' \geq_{\mathrm{d}} X$ where $\geq_{\mathrm{d}}$ means greater or equal in distribution.

Lemma 2 (Stochastic Order)**.** *Consider sets of independent geometric random variables $X'(n)$, $Y'(n)$ with parameter p_R and $Z'(n,j)$ with parameter p_P for $n \geq 1$ and $j \in [0, w-1]$. Use X', Y', and Z' to define $S'_q(m,n)$ and $S'_r(m,n)$ in accordance with Lem. 1. It holds for $n \geq m \geq 1$ that:*

i) $S'_q(m,n) \geq_d S_q(m,n)$,

ii) $S'_r(m,n) \geq_d S_r(m,n)$.

Proof. i) The success of each packet transmission including the acknowledgement is an independent Bernoulli trial with parameter p_R. If $w = 1$, it follows that starting at $D_q(n) + 1$ it takes $X(n+1)$ retransmissions of packet n before service is available for packet $n+1$ where $\mathsf{P}[X(n+1) = x] = (1-p_R)^x p_R$ for $x \geq 0$ and $n \geq 1$, i.e., the $X(n)$ are independent and identically distributed (iid) geometric random variables.

Otherwise, if $w > 1$, the history of the $w-1$ time-slots preceding $D_q(n)$ has to be considered. It holds that

$$\mathsf{P}[X(n+1) = 0 | I(D_q(n) - w + 1) = 0] = 1. \tag{5}$$

The reason is that since time-slot $D_q(n) - w + 1$ was not used, there cannot be a timeout in time-slot $D_q(n)+1$ so that time-slot $D_q(n)+1$ is generally available for transmission of a new packet and hence $X(n+1) = 0$. Similarly, it holds that

$$\mathsf{P}[X(n+1) = x | I(D_q(n)-w+1) = 1, I(D_q(n)-w+2) = 0] = \begin{cases} p_R & \text{, for } x = 0, \\ (1-p_R) & \text{, for } x = 1, \end{cases} \tag{6}$$

assuming $w > 2$. In this case, the transmission in time-slot $D_q(n)-w+1$ may cause a retransmission in time-slot $D_q(n)+1$ with probability $(1-p_R)$, whereas time-slot $D_q(n)+2$ is generally available since time-slot $D_q(n) - w + 2$ was not used. Following the same argument, we eventually obtain

$$\mathsf{P}[X(n+1) = x | I(D_q(n)-w+1) = 1, \ldots, I(D_q(n)-2) = 1, I(D_q(n)-1) = 0] = \begin{cases} (1-p_R)^x p_R & \text{, for } x \leq w-3, \\ (1-p_R)^x & \text{, for } x = w-2, \end{cases} \tag{7}$$

assuming $w > 3$. On the other hand, if all $w-1$ time-slots preceding $D_q(n)$ have been used it holds for $x \geq 0$ that

$$\mathsf{P}[X(n+1) = x | I(D_q(n)-w+1) = 1, \ldots, I(D_q(n)-1) = 1] = (1-p_R)^x p_R. \tag{8}$$

Clearly, the $X(n)$ in Eq. (8) are iid geometric random variables with parameter p_R, while the $X(n)$ in Eqs. (5)-(7) are truncated geometric random variables with the same parameter p_R, including the degenerate case in Eq. (5). Since the individual packet transmissions are independent, it follows that $\sum_{\nu=m+1}^{n} X'(\nu) \geq_{\mathrm{d}} \sum_{\nu=m+1}^{n} X(\nu)$ for $n \geq m \geq 1$. The same argument applies to show that $Y(n)$ is a truncated geometric random variable, and hence $S'_q(m,n) \geq_{\mathrm{d}} S_q(m,n)$.

ii) Each transmission of a packet is an independent Bernoulli trial with probability p_P. Hence, for any packet that is transmitted in time-slot $D_q(n) - j$, i.e., $I(D_q(n) - j) = 1$, the number of retransmissions $Z(n,j)$ are iid geometric random variables with parameter p_P for $j \in [0, w-1]$. On the other

hand, if $I(D_q(n)-j)=0$ we have $Z(n,j)=0$. Hence $Z'(n,j) \geq_{\mathrm{d}} Z(n,j)$ for $n \geq 1$ and $j \in [0, w-1]$ and $S'_r(m,n) \geq_{\mathrm{d}} S_r(m,n)$ for $n \geq m \geq 1$. □

III. Delay Bounds for General Arrivals

We use the stochastic network calculus, that is a set of methods for derivation of statistical performance bounds see, e.g., [3], [24]–[29], to analyze the waiting time and sojourn time of the ARQ protocol. Specifically, we use methods that operate on moment generating functions (MGFs) of the traffic and service, respectively. Different from the standard space-domain approach of the network calculus that is expressed in a min-plus algebra, we use the definition of max-plus server. Statistical performance bounds for max-plus servers are also provided by [21], [23].

A. Statistical Traffic and Service Characterization

We characterize the data traffic and the service by parameterized envelopes of their MGFs. The model is based on the (σ,ρ)-envelope of [3]. The MGF of a random variable X is defined as $\mathsf{M}_X(\theta)=\mathsf{E}[e^{\theta X}]$ where θ is a free parameter.

Definition 2 (Traffic and Service Parameters)**.** *An arrival process $A(m,n)$ is (σ_A,ρ_A)-lower constrained if for all $n \geq m \geq 1$ and $\theta > 0$ it holds that*

$$\mathsf{E}\left[e^{-\theta A(m,n)}\right] \leq e^{-\theta(\rho_A(-\theta)(n-m)-\sigma_A(-\theta))}.$$

A service process $S(m,n)$ is (σ_S,ρ_S)-upper constrained if for all $n \geq m \geq 1$ and $\theta > 0$ it holds that

$$\mathsf{E}\left[e^{\theta S(m,n)}\right] \leq e^{\theta(\rho_S(\theta)(n-m)+\sigma_S(\theta))}.$$

The parameters (σ,ρ) can be thought of as an effective rate ρ and a burstiness term σ. Beyond the existence of a bounded MGF, Def. 2 does not make any assumptions about the distribution of the arrivals and the service. Since the MGFs of a wide variety of distributions are known, the (σ,ρ) parameters of different arrival and service processes are readily available. For the purpose of our evaluation of ARQ protocols, we will simply use arrivals with iid geometric inter-arrival times with parameter p_A, i.e., $\mathsf{P}[A(n-1,n)=a]=(1-p_A)^a p_A$ for $a \geq 0$ and $n \geq 1$. It follows that $A(n)$ is negative binomial and $\mathsf{E}[e^{-\theta A(m,n)}]=(p_A/(1-(1-p_A)e^{-\theta}))^{n-m}$ so that for $\theta > 0$

$$\rho_A(-\theta)=-\frac{1}{\theta}\ln\left(\frac{p_A}{1-(1-p_A)e^{-\theta}}\right),$$

and $\sigma_A(-\theta)=0$. In the case of arrivals with independent increments it holds generally that $\sigma_A(-\theta)=0$.

B. Statistical Delay Bounds

The following theorem derives statistical bounds of the waiting time and the sojourn time of the ARQ protocol for general traffic arrivals with parameters defined by Def. 2. In the following T denotes delay that is either the waiting time defined as $T_q(n)=D_q(n)-A(n)$ or the sojourn time $T_r(n)=D_r(n)-A(n)$ for $n \geq 1$, depending on the service parameters that are used.

Theorem 1 (Delay Bounds for General Arrivals)**.** *Consider arrivals with parameters $(\sigma_A(-\theta),\rho_A(-\theta))$ that are transmitted by the ARQ protocol. It holds for the delay for $n \geq 1$ and $\tau \geq 0$ that*

$$\mathsf{P}[T(n) > \tau] \leq \alpha e^{-\theta\tau},$$

where

$$\alpha=\frac{e^{\theta(\sigma_A(-\theta)+\sigma_S(\theta))}}{1-e^{-\theta(\rho_A(-\theta)-\rho_S(\theta))}}.$$

The free parameter $\theta > 0$ has to satisfy $\rho_S(\theta) < \rho_A(-\theta)$, where

$$\rho_S(\theta)=1+\frac{1}{\theta}\ln\left(\frac{p_R}{1-(1-p_R)e^{\theta}}\right),$$

for $\theta \leq -\ln(1-p_R)$. Above, T is the waiting time if

$$\sigma_S(\theta)=\rho_S(\theta)-1,$$

or the sojourn time if

$$\sigma_S(\theta)=\rho_S(\theta)+t_P+\frac{1}{\theta}\ln\left(\frac{p_P}{1-(1-p_P)e^{\theta t_O}}\frac{1-e^{-\theta w}}{1-e^{-\theta}}\right),$$

for $\theta < -\ln(1-p_P)/t_O$.

The free parameter θ in Th. 1 can be optimized to obtain the smallest delay bound.

Proof. First, we derive the parameters of the service processes. For geometric random variables $X'(n)$ and $Y'(n)$ with parameter p_R, we have $\mathsf{M}_{X'}(\theta)=\mathsf{M}_{Y'}(\theta)=p_R/(1-(1-p_R)e^{\theta})$ for $\theta < -\ln(1-p_R)$. The MGF of $S'_q(m,n)$ in Lem. 2 for $\theta \leq -\ln(1-p_R)$ is

$$\mathsf{E}\left[e^{\theta S'_q(m,n)}\right] \leq \frac{p_R}{1-(1-p_R)e^{\theta}}\left(\frac{p_R e^{\theta}}{1-(1-p_R)e^{\theta}}\right)^{n-m},$$

where we estimated $\min\{w-1, Y(m)\} \leq Y(m)$. With $\mathsf{M}_{Z't_O}(\theta)=p_P/(1-(1-p_P)e^{\theta t_O})$ for $\theta < -\ln(1-p_P)/t_O$, the MGF of $S'_r(m,n)$ is

$$\mathsf{E}\left[e^{\theta S'_r(m,n)}\right] \leq \mathsf{E}\left[e^{\theta S'_q(m,n)}\right]\frac{p_P e^{\theta(t_P+1)}}{1-(1-p_P)e^{\theta t_O}}\frac{1-e^{-\theta w}}{1-e^{-\theta}},$$

for $\theta \in [0, -\ln(1-p_P)/t_O)$. Above we used that $X'(n)$, $Y'(n)$, and $Z'(n,j)$ are independent geometric random variables and

$$\mathsf{E}\left[e^{\theta \max_{j\in[0,w-1]}\{Z'(n,j)t_O-j\}}\right] \leq \sum_{j=0}^{w-1}\mathsf{E}\left[e^{\theta(Z'(n,j)t_O-j)}\right]$$

for $\theta \geq 0$, where we finally solved the sum of the geometric series. With $S(m,n) \leq_d S'(m,n)$ it follows for $\theta \geq 0$ that $\mathsf{E}[e^{\theta S(m,n)}] \leq \mathsf{E}[e^{\theta S'(m,n)}]$ and the parameters of the service processes are obtained directly.

The rest of the proof uses basic steps of the stochastic network calculus as in [23]. By insertion of the definition of max-plus server into the definition of delay $T(n)=D(n)-A(n)$ we have $T(n) \leq \max_{m\in[1,n]}\{S(m,n)-A(m,n)\}$. We obtain for $\theta \geq 0$ that

$$\mathsf{E}\left[e^{\theta T(n)}\right] \leq \sum_{m=1}^{n}\mathsf{E}\left[e^{\theta S(m,n)}\right]\mathsf{E}\left[e^{-\theta A(m,n)}\right],$$

where we estimated the maximum by the sum of its arguments and used the statistical independence of arrivals and service. By insertion of the (σ, ρ) parameters we have

$$\mathsf{E}\left[e^{\theta T(n)}\right] \leq e^{\theta(\sigma_A(-\theta)+\sigma_S(\theta))} \sum_{m=1}^{n} e^{-\theta(\rho_A(-\theta)-\rho_S(\theta))(n-m)}.$$

Next, we estimate

$$\sum_{m=1}^{n} e^{-\theta(\rho_A(-\theta)-\rho_S(\theta))(n-m)} \leq \sum_{\nu=0}^{\infty} \left(e^{-\theta(\rho_A(-\theta)-\rho_S(\theta))}\right)^{\nu} = \frac{1}{1-e^{-\theta(\rho_A(-\theta)-\rho_S(\theta))}},$$

for $\rho_S(\theta) < \rho_A(-\theta)$. Applying Chernoff's bound $\mathsf{P}[T(n) \geq \tau] \leq e^{-\theta\tau}\mathsf{E}\left[e^{\theta T(n)}\right]$ for $\theta \geq 0$ completes the proof. □

C. Sojourn Time Bound via the Waiting Time Bound

Compared to the waiting time bound, the sojourn time bound in Th. 1 has an additional constraint on the range of the parameter θ. This constraint can be avoided by the following derivation of the sojourn time that can provide numerically tighter bounds. With Lem. 1 the sojourn time can be expressed by the waiting time as $T_r(n) \leq T_q(n) + t_P + 1 + \max_{j\in[0,w-1]}\{Z(n,j)t_O - j\}$. Hence, with Lem. 2 we can estimate

$$T_r(n) \leq T_q(n) + t_P + 1 + \max_{j\in[0,w-1]}\{Z'(n,j)t_O\}. \quad (9)$$

Since the $Z'(n,j)$ are iid geometric random variables with parameter p_P we have $\mathsf{P}[Z'(n,j) > z] = (1-p_P)^{z+1}$ for $z \geq 0$. Hence, for $\kappa \in \{0, t_O, 2t_O, \dots\}$

$$\mathsf{P}\left[\max_{j\in[0,w-1]}\{Z'(n,j)\} \leq \frac{\kappa}{t_O}\right] = \left(1-(1-p_P)^{\frac{\kappa}{t_O}+1}\right)^{w}. \quad (10)$$

With Th. 1 it holds that $\mathsf{P}[T_q(n) \leq \tau] \geq 1-\min\{1, \alpha e^{-\theta\tau}\} = [1-\alpha e^{-\theta\tau}]^{+}$ where $[x]^{+} = \max\{0, x\}$. Using the independence of $T_q(n)$ and $Z'(n,j)$, we have from Eq. (9) that

$$\mathsf{P}[T_r(n) \leq \tau] \geq \sum_{\kappa\in\left\{0,t_O,2t_O,\dots,\left\lfloor\frac{\tau-t_P-1}{t_O}\right\rfloor t_O\right\}} \mathsf{P}[T_q(n) \leq \tau - t_P - 1 - \kappa]\ \mathsf{P}\left[\max_{j\in[0,w-1]}\{Z'(n,j)\} = \frac{\kappa}{t_O}\right],$$

for $\tau \geq t_P + 1$. By insertion of $T_q(n)$ from Th. 1 and Eq. (10) it follows that

$$\mathsf{P}[T_r(n) > \tau] \leq 1 - \sum_{\kappa\in\left\{0,t_O,2t_O,\dots,\left\lfloor\frac{\tau-t_P-1}{t_O}\right\rfloor t_O\right\}} \left[1-\alpha e^{-\theta(\tau-t_P-1-\kappa)}\right]^{+} \left(\left(1-(1-p_P)^{\frac{\kappa}{t_O}+1}\right)^{w} - \left(1-(1-p_P)^{\frac{\kappa}{t_O}}\right)^{w}\right), \quad (11)$$

for $\tau \geq t_P + 1$. We use Eq. (11) for numerical evaluation.

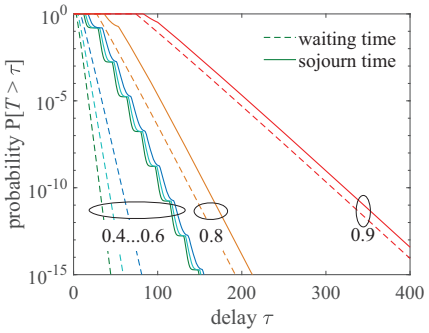

Fig. 4. Tail decay of the waiting time bound (dashed lines) and the sojourn time bound (solid lines). The curves are shown for utilizations of $\{0.4, 0.5, 0.6, 0.8, 0.9\}$.

IV. Performance Evaluation

In this section, we study the impact of the different parameters of the ARQ service process and show a comparison with the simpler cases of stop-and-wait ARQ and ARQ with instantaneous feedback, respectively, that are frequently used in the literature. Since our focus is on the service of the ARQ protocol, we only show results for a simple traffic model with geometric inter-arrival times with parameter p_A. Models of more general arrival processes are readily available in the literature, e.g., in [23], [29].

A. Parameter Study

Speed of the tail decay: First, we consider the tail distribution of the waiting time bound and the sojourn time bound. The waiting time bound is computed using Th. 1 for a range of parameters θ where we finally select the smallest waiting time bound. The sojourn time bound is computed from the waiting time bound using Eq. (11). Fig. 4 shows the results for a propagation delay $t_P = 8$, probability of successful packet transmission $p_P = 0.99$ and successful acknowledgement transmission $p_K = 1$ so that $p_R = 0.99$. The utilization is varied in $\{0.4, 0.5, 0.6, 0.8, 0.9\}$. The utilization is defined as the quotient of the mean arrival rate $p_A/(1-p_A)$ and the probability of successful transmission p_R. Generally, we observe that the waiting time bound and the sojourn time bound increase with increasing utilization.

For low to moderate utilizations up to 0.6, retransmissions account for the largest part of the sojourn time. The effect of the retransmissions, that occur with probability 10^{-2} and take one round-trip-time $t_R = 17$, is visible in the stepped shape of the sojourn time curves. Although the waiting time increases with increasing utilization, the sojourn time is only marginally influenced by the utilization. This is due to the fact that the tail distribution of the waiting time decays faster than the tail distribution of the sojourn time.

In case of high utilization of 0.8 and 0.9, respectively, the effect is reversed and the waiting time dominates the sojourn time. This is a consequence of the speed of the tail decay of the waiting time bound that decreases with increasing utilization and eventually becomes slower than the tail decay that is due to the retransmissions.

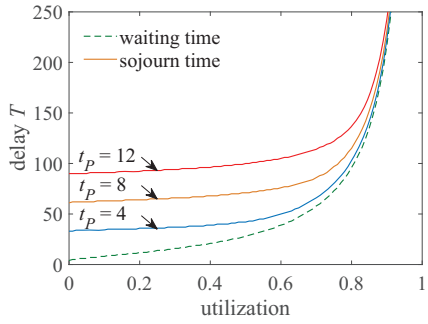

Fig. 5. Waiting time bounds (dashed lines) and sojourn time bounds (solid lines) as functions of the utilization. The bounds are exceeded at most with probability $\varepsilon = 10^{-6}$. The propagation delay is $t_P \in \{4, 8, 12\}$. The propagation delay has no impact on the waiting time bound.

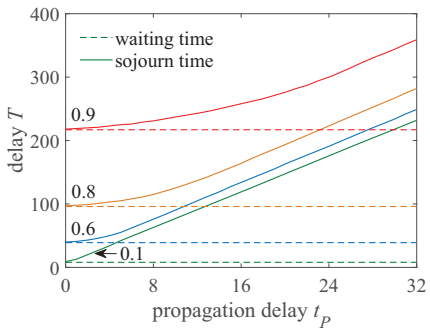

Fig. 6. Waiting time bounds (dashed lines) and sojourn time bounds (solid lines) as functions of the propagation delay t_P. The utilization is varied in $\{0.1, 0.6, 0.8, 0.9\}$.

The same effect is also visible in Fig. 5 that shows the waiting time bound and the sojourn time bound as functions of the utilization. The bounds are exceeded at most with probability $\varepsilon = 10^{-6}$ and the parameters are $p_P = 0.99$, $p_K = 1$, and $t_P \in \{4, 8, 12\}$. The sojourn time is mostly due to retransmissions and only marginally influenced by the utilization, as long as the utilization is moderate. In contrast, the waiting time increases with the utilization and eventually dominates the sojourn time in case of high utilization.

Impact of the propagation delay: Fig. 5 also shows the effect of the propagation delay. While the waiting time bound does not depend on the propagation delay, the sojourn time bound shows a linear increase with the propagation delay. To see how the propagation delay influences the sojourn time consider the case $t_P = 8$ that corresponds to a round-trip-time of $t_R = 17$. Given the packet loss probability of 10^{-2}, a packet requires 3 retransmissions with probability 10^{-6} resulting in a delay of $3t_R + t_P + 1 = 60$ not including waiting times.

The relation between the sojourn time and the propagation delay is also depicted in Fig. 6. The parameters are $\varepsilon = 10^{-6}$, $p_P = 0.99$, $p_K = 1$, and the utilization is in $\{0.1, 0.6, 0.8, 0.9\}$. As before, the waiting time bound is independent of t_P and increases progressively with the utilization. For $t_P = 0$, i.e., in case of instantaneous feedback, the sojourn time bound closely approaches the waiting time bound. With increasing t_P the sojourn time bound eventually shows a linear growth that is due to retransmission delays. The reason is that the tail distribution of the retransmission delays decays slower with increasing t_P so that the retransmission delays eventually dominate the waiting time (more quickly in case of lower utilization). As a consequence, the utilization has decreasing effect on the sojourn time bound, visible by the gradual convergence of the curves in Fig. 6 for large t_P.

Impact of the loss probability: Fig. 7 shows the influence of the loss probability. Since the capacity of the ARQ protocol decreases with increasing loss probability, we also reduce the traffic arrival rate accordingly to keep the utilization fixed. We use parameters $\varepsilon = 10^{-6}$, $p_K = 1$, $t_P = 8$, and the utilization is in $\{0.1, 0.6, 0.7, 0.8\}$. For a fixed utilization, we observe only a minor increase of the waiting time bound with the loss probability $1 - p_R$. The sojourn time bound on the other hand shows a significant impact of the loss probability. Again, in case of low utilization, the waiting time is quickly dominated by the retransmission delays. The steps of the sojourn time curve reflect that with increasing loss probability additional retransmissions that take $t_R = 17$ each are considered by the statistical bound.

B. Special Cases

For comparison, we also include models of the special cases of ARQ with instantaneous feedback and stop-and-wait ARQ, respectively. These two cases are frequently considered in the literature due to their simplicity.

1) Instantaneous Feedback: In case of instantaneous feedback, i.e., $t_P = 0$, the round-trip-time is $t_R = 1$. Hence, a timeout $t_O = 1$ and a window $w = 1$ are sufficient. In this case Lem. 1 gives $S_q(m,n) = n - m + \sum_{\nu=m+1}^{n} X(\nu)$ and

$$S_r(m,n) = n - m + 1 + \sum_{\nu=m+1}^{n} X(\nu) + Z(n,0).$$

With Lem. 2 the MGF of the service process is estimated as

$$\mathsf{E}\left[e^{\theta S'_r(m,n)}\right] \leq \left(\frac{p_R e^{\theta}}{1-(1-p_R)e^{\theta}}\right)^{n-m} \frac{p_P e^{\theta}}{1-(1-p_P)e^{\theta}}.$$

2) Stop-and-Wait ARQ: For the stop-and-wait ARQ protocol, we consider the propagation delay $t_P \geq 0$ and the round-trip-time is $t_R = 1 + 2t_P$. The timeout is selected to be $t_O = \lceil t_R \rceil$, however, the window is due to the stop-and-wait protocol $w = 1$, generally. Also, due to the stop-and-wait protocol, each packet transmission or retransmission blocks the channel for t_O units of time. The service process follows as $S_q(m,n) = t_O\left(n - m + \sum_{\nu=m+1}^{n} X(\nu)\right)$ and

$$S_r(m,n) = t_O\left(n - m + \sum_{\nu=m+1}^{n} X(\nu) + Z(n,0)\right) + t_P + 1.$$

With Lem. 2 the MGF of the service process is estimated as

$$\mathsf{E}\left[e^{\theta S'_r(m,n)}\right] \leq \left(\frac{p_R e^{\theta t_O}}{1-(1-p_R)e^{\theta t_O}}\right)^{n-m} \frac{p_P e^{\theta(t_P+1)}}{1-(1-p_P)e^{\theta t_O}}.$$

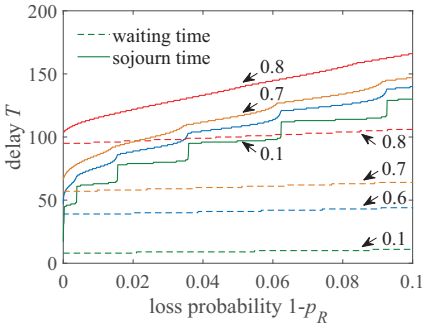

Fig. 7. Waiting time bounds (dashed lines) and sojourn time bounds (solid lines) as functions of the packet loss probability $1-p_R$. The utilization is varied in $\{0.1, 0.6, 0.7, 0.8\}$.

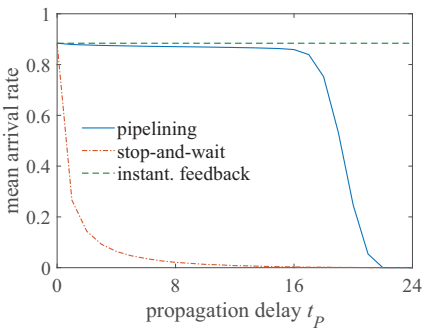

Fig. 8. Mean arrival rate up to which the protocols, ARQ with pipelining and selective repeat, stop-and-wait ARQ, and ARQ with instantaneous feedback, can ensure a sojourn time bound of $T_r = 200$ that is exceeded at most with probability $\varepsilon = 10^{-6}$.

Fig. 8 shows a comparison of the three cases: ARQ with pipelining and selective repeat, stop-and-wait ARQ, and ARQ with instantaneous feedback. We consider a sojourn time bound of $T_r = 200$ that must not be exceeded with a probability of more than $\varepsilon = 10^{-6}$. We depict the largest mean arrival rate for which the ARQ protocol can ensure this sojourn time bound. The parameters are $p_P = 0.99$ and $p_K = 1$. For $t_P = 0$ all three protocols are identical. They differ, however, when t_P is increased. Under the assumption of instantaneous feedback, the propagation delay is neglected and hence it does not show any impact. In contrast, the throughput of stop-and-wait ARQ reduces quickly if the propagation delay is increased. This problem is resolved by pipelining that can sustain high mean arrival rates also in the case of non-negligible t_P. Eventually, when t_P exceeds a critical value (approximately 20) retransmission delays make it impossible to ensure the sojourn time bound and the sustainable mean arrival rate drops quickly to zero.

V. Arrivals with Independent Increments

We also include delay bounds that take advantage of arrivals with iid increments. After stating the result, we provide a comparison with simulation results that shows the good accuracy of the bounds.

Theorem 2 (Delay Bounds for Arrivals with IID Increments)**.** *Consider arrivals with iid inter-arrival times and parameter $\rho_A(-\theta)$ that are transmitted by the ARQ protocol. It holds for the delay for $n \geq 1$ and $\tau \geq 0$ that*

$$\mathsf{P}[T(n) > \tau] \leq \alpha e^{-\theta\tau},$$

where

$$\alpha = e^{\theta\sigma_S(\theta)}.$$

The free parameter $\theta > 0$ has to satisfy $\rho_S(\theta) \leq \rho_A(-\theta)$, where $\rho_S(\theta)$ is given in Th. 1. Above, T is the waiting time if

$$\sigma_S(\theta) = w - 1,$$

or the sojourn time if

$$\sigma_S(\theta) = w + t_P + \frac{1}{\theta}\ln\left(\frac{p_P}{1-(1-p_P)e^{\theta t_O}}\frac{1-e^{-\theta w}}{1-e^{-\theta}}\right),$$

for $\theta < -\ln(1-p_P)/t_O$.

Proof. The proof uses Doob's martingale inequality [30] as, e.g., in [23], [27], [31], [32]. The parameters of the service processes are derived as in the proof of Th. 1, where we now estimate $\min\{w-1, Y(m)\}$ by $w-1$ instead. By insertion of the definition of max-plus server into the definition of delay $T(n) = D(n) - A(n)$ and using the ARQ service process from Lem. 1 we have

$$T(n) \leq \max_{m\in[1,n]}\left\{V(n)+m-1+\sum_{\nu=n-m+2}^{n} X(\nu) - A(n-m+1,n)\right\},$$

where $V(n) = w-1$ in the case of the waiting time and $V(n) = w + t_P + \max_{j\in[0,w-1]}\{Z(n,j)t_O - j\}$ in the case of the sojourn time, respectively. For $\theta > 0$ and with Lem. 2 and defining $V'(n)$ accordingly we can write

$$\mathsf{P}[T(n) > \tau] \leq \mathsf{P}\Bigg[\max_{m\in[1,n]} \left\{e^{\theta(V'(n)+m-1+\sum_{\nu=n-m+2}^{n} X'(\nu) - A(n-m+1,n))}\right\} > e^{\theta\tau}\Bigg].$$

Now consider the increment process

$$U(m) = e^{\theta(V'(n)+m-1+\sum_{\nu=n-m+2}^{n} X'(\nu) - A(n-m+1,n))}.$$

Hence, $U(m+1) = U(m)e^{\theta(1+X'(n-m+1)-A(n-m,n-m+1))}$ that has the conditional expectation

$$\begin{aligned}&\mathsf{E}[U(m+1) \mid U(m), U(m-1), \ldots, U(1)]\\ &\quad = U(m)e^{\theta}\mathsf{E}\left[e^{\theta X'(n-m+1)}\right]\mathsf{E}\left[e^{-\theta A(n-m,n-m+1)}\right],\end{aligned}$$

where we used the independence of X' and A. If $\rho_S(\theta) \leq \rho_A(-\theta)$ it follows with $e^{\theta}\mathsf{E}[e^{\theta X'(n-m+1)}] \leq e^{\theta\rho_S(\theta)}$ and $\mathsf{E}[e^{-\theta A(n-m,n-m+1)}] \leq e^{-\theta\rho_A(-\theta)}$ that

$$e^{\theta}\mathsf{E}[e^{\theta X'(n-m+1)}]\mathsf{E}[e^{-\theta A(n-m,n-m+1)}] \leq 1. \qquad (12)$$

Thus, $\mathsf{E}[U(m+1) \mid U(m), U(m-1), \ldots, U(1)] \leq U(m)$, so that $U(m)$ is a supermartingale. With Doob's inequality [30] it holds that [32]

$$x\mathsf{P}\left[\max_{m\in[1,n]}\{U(m)\} \geq x\right] \leq \mathsf{E}[U(1)].$$

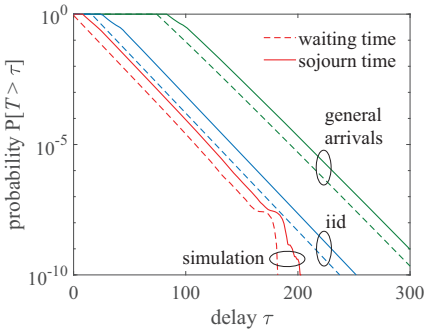

Fig. 9. Comparison of Th. 1 for general arrivals and Th. 2 for arrivals with iid increments with simulation results. In all cases the arrivals have iid geometric inter-arrival times.

With $\mathsf{E}[U(1)] = \mathsf{E}[e^{\theta V'(n)}] \leq e^{\theta \sigma_S(\theta)}$ and letting $x = e^{\theta\tau}$ we have $\mathsf{P}[T(n) \geq \tau] \leq e^{\theta\sigma_S(\theta)} e^{-\theta\tau}$. □

In Fig. 9, we compare the waiting time bounds and sojourn time bounds from Th. 1 and Th. 2 with simulation results. In both cases, we derived the sojourn time bound from the waiting time bound using Eq. (11). The simulation results comprise 10^{10} samples. The parameters are $p_P = 0.99$, $p_K = 1$, $t_P = 8$ and the utilization is 0.9. While Th. 2 provides tighter bounds for arrivals with iid increments, Th. 1 also shows the correct speed of the tail decay. Further Th. 1 is applicable also if the arrival increments are not iid.

VI. Conclusion

We modelled the service process of a general class of ARQ protocols with pipelining and selective repeat and derived statistical bounds of the waiting time and the sojourn time of the ARQ system for a general packet arrival process. We found two regimes, where the sojourn time is dominated either by the waiting time or by retransmission delays, depending on whichever has the slower decay rate of the tail distribution. We showed that the decay rate of the waiting time bound depends primarily on the utilization of the ARQ system, whereas the decay rate of the retransmission delays is mostly determined by the round-trip-time and the packet loss probability. An important conclusion is that in networks with non-negligible round-trip-times that operate under the second regime, actions that reduce the utilization can hardly remedy large sojourn times. Since retransmission delays become insignificant in the case of small round-trip-times, known ARQ models that assume instantaneous feedback reveal only the utilization-dependent waiting time. Besides such special cases, we also included performance bounds for arrivals with iid increments and showed the accuracy compared to simulation results.

References

[1] L. H. Ozarow, S. Shamai, and A. D. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Veh. Technol.*, vol. 43, no. 2, pp. 359–378, May 1994.

[2] S. V. Hanly and D. N. Tse, "Multiaccess fading channels-part ii: Delay-limited capacities," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2816–2831, Nov. 1998.

[3] C.-S. Chang, *Performance Guarantees in Communication Networks*. Springer-Verlag, 2000.

[4] D. O. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[5] R. L. Cruz, "Quality of service management in Integrated Services networks," in *Proc. of Semi-Annual Research Review, Center of Wireless Communication, UCSD*, Jun. 1996.

[6] Y. Jiang and P. J. Emstad, "Analysis of stochastic service guarantees in communication networks: A server model," in *Proc. of IWQoS*, Jun. 2005, pp. 233–245.

[7] M. Fidler, "A network calculus approach to probabilistic quality of service analysis of fading channels," in *IEEE Globecom*, Nov. 2006.

[8] C. Li, H. Che, and S. Li, "A wireless channel capacity model for quality of service," *IEEE Trans. Wireless Commun.*, vol. 6, no. 1, pp. 356–366, 2007.

[9] H. Al-Zubaidy, J. Liebeherr, and A. Burchard, "A (min,x)-network calculus for multi-hop fading channels." in *IEEE INFOCOM*, Apr. 2013.

[10] M. Fidler, R. Lübben, and N. Becker, "Capacity-Delay-Error-Boundaries: A Composable Model of Sources and Systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1280–1294, Mar. 2015.

[11] S. Schiessl, J. Gross, and H. Al-Zubaidy, "Delay analysis for wireless fading channels with finite blocklength channel conding," in *Proc. of ACM MSWiM*, Nov. 2015.

[12] S. Akin and M. Fidler, "Backlog and delay reasoning in HARQ systems," in *Proc. of ITC 27*, Sep. 2015.

[13] S. Schiessl, F. Naghibi, H. Al-Zubaidy, M. Fidler, and J. Gross, "On the delay performance of interference channels," in *Proc. of IFIP Networking*, May 2016.

[14] S. Akin and M. C. Gursoy, "Effective capacity analysis of cognitive radio channels for quality of service provisioning," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3354–3364, Nov. 2010.

[15] S. Akin and M. Fidler, "On the transmission rate strategies in cognitive radios," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 2335–2350, Mar. 2016.

[16] K. Mahmood, A. Rizk, and Y. Jiang, "On the flow-level delay of a spatial multiplexing MIMO wireless channel," in *IEEE ICC*, Jun. 2011.

[17] F. Ciucu, "On the scaling of non-asymptotic capacity in multi-access networks with bursty traffic." in *Proc. of IEEE ISIT*, Aug. 2011.

[18] H. Wang, J. Schmitt, and F. Ciucu, "Performance modelling and analysis of unreliable links with retransmissions using network calculus," in *Proc. of ITC 25*, Sep. 2013.

[19] J.-Y. Le Boudec and P. Thiran, *Network Calculus A Theory of Deterministic Queuing Systems for the Internet*. Springer-Verlag, 2001.

[20] J. Liebeherr, *Duality of the Max-Plus and Min-Plus Network Calculus*. Now Publishers, 2017.

[21] J. Xie and Y. Jiang, "Stochastic network calculus models under max-plus algebra," in *Proc. of IEEE GLOBECOM*, 2009.

[22] R. Lübben, M. Fidler, and J. Liebeherr, "Stochastic bandwidth estimation in networks with random service," *IEEE/ACM Trans. Netw.*, vol. 22, no. 2, pp. 484–497, Apr. 2014.

[23] M. Fidler, B. Walker, and Y. Jiang, "Non-Asymptotic Delay Bounds for Multi-Server Systems with Synchronization Constraints," *IEEE Trans. Parallel Distrib. Syst.*, Jan. 2018.

[24] F. Ciucu, A. Burchard, and J. Liebeherr, "Scaling properties of statistical end-to-end bounds in the network calculus," *IEEE/ACM Trans. Netw.*, vol. 14, no. 6, pp. 2300–2312, Jun. 2006.

[25] M. Fidler, "An end-to-end probabilistic network calculus with moment generating functions," in *Proc. of IWQoS*, Jun. 2006, pp. 261–270.

[26] Y. Jiang and Y. Liu, *Stochastic Network Calculus*. Springer-Verlag, Sep. 2008.

[27] F. Ciucu and J. Schmitt, "Perspectives on network calculus - no free lunch but still good value," in *Proc. of ACM SIGCOMM*, Aug. 2012, pp. 311–322.

[28] M. Fidler, "A survey of deterministic and stochastic service curve models in the network calculus," *IEEE Commun. Surveys Tuts.*, vol. 12, no. 1, pp. 59–86, 2010.

[29] M. Fidler and A. Rizk, "A guide to the stochastic network calculus," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 92–105, Mar. 2015.

[30] J. L. Doob, *Stochastic Processes*. Wiley, 1953.

[31] J. F. C. Kingman, "A martingale inequality in the theory of queues," *Math. Proc. Cambridge*, vol. 60, no. 2, pp. 359–361, Apr. 1964.

[32] Y. Jiang, "Network calculus and queueing theory: Two sides of one coin," in *Proc. of VALUETOOLS*, 2009, pp. 1–12, Invited Paper.