Principle and Limit for Guaranteeing the Mean Access Delay in S-ALOHA

Zhenyu Cao, Hu Jin, *Senior Member, IEEE*, Swades De, *Senior Member, IEEE*, and Jun-Bae Seo, *Member, IEEE*

***Abstract*—The access delay of Slotted ALOHA (S-ALOHA) is intricately coupled with backoff intervals and the number of backlogged users, while new arrivals continuously join the system according to the input rate. Consequently, guaranteeing a mean access delay constraint is a challenging problem since the system must control these interdependent variables. This work proposes an optimization framework that determines the maximum allowable input rate to satisfy the access delay constraint, where a throughput-maximizing backoff algorithm based on the extended Kalman filter (EKF) is developed. To solve the optimization problem, extensive analysis of S-ALOHA is conducted with access delay characterization. The results demonstrate that while the input rate is constrained through the proposed framework, the proposed EKF-based backoff algorithm successfully guarantees the access delay without the knowledge of the exact backlog size.**

***Index Terms*—S-ALOHA, Access delay, extended Kalman filter**

I. Introduction

WHILE the throughput performance of Slotted ALOHA (S-ALOHA) has been extensively studied, the analysis of access delay remains comparatively limited. The primary challenge arises from the interdependence between the backoff algorithm and the system input rate, where the latter determines the influx of newly contending packets. This coupling creates a feedback loop: The input rate continuously increases contention, while the backoff algorithm determines the cumulative random delays experienced prior to successful transmission. Access delay analysis thus requires modeling the interaction between arrivals and retransmission mechanisms, making it significantly more challenging than throughput analysis. This letter addresses this gap by developing a tractable analytical framework for guaranteeing the mean access delay of S-ALOHA systems.

Despite this complexity, several studies have characterized the access delay, which can be divided into two groups based on whether users maintain queues to store incoming packets. For systems with packet queuing [1]–[3], additional queueing delay must be considered, significantly complicating the analysis. For users with a single packet to send, which is our assumption in this work, Tobagi [4] derived the Z-transform of the access delay distribution in slotted ALOHA when users retransmit with probability p. Yang and Yum [5] obtained delay distributions under various retransmission policies, including uniform backoff (UB), geometric backoff (GB), and binary exponential backoff (BEB). The delay properties of BEB were further analyzed in [6], [7], where conditions for stability and finite mean delay were established. In [8], it was demonstrated that the access delay distributions of GB, BEB, and UB can be approximated by Gaussian distributions. For multichannel slotted ALOHA, the access delay distributions under GB and UB algorithms were obtained in [9], motivated by ultra-reliable low-latency communications (URLLC). In [10], the Z-transform of access delay with GB was derived, and the access probability was optimized to balance sum-rate and delay performance. It is notable that several analyses [6], [7], [10], assume a saturated condition with N users each always having a packet to send, which simplifies user interactions, enabling tractable analysis, but remains an approximation.

In contrast with prior studies focused on specific backoff schemes such as GB, UB, or BEB, this letter formulates an optimization framework that guarantees a prescribed mean access delay for S-ALOHA, where the retransmission probability is adaptively controlled as a function of the backlog size. Following [4], [5], [9], we adopt the unsaturated single-packet model, relevant to sporadic traffic typical of IoT sensor networks and low-duty-cycle transmissions. A key observation is that throughput maximization and delay minimization are inherently equivalent, and guaranteeing a delay constraint requires limiting the input rate under a given backoff algorithm.

The main contribution of this letter is a framework that determines the maximum input rate while guaranteeing a specified delay limit. We first consider a genie-aided system with perfect backlog knowledge as a theoretical benchmark. To make this practical, we develop a backlog estimation algorithm using the extended Kalman filter (EKF) (Sec. II-C). Our optimization framework explicitly accounts for estimation errors to determine the maximum achievable input rate that maintains the delay guarantee. Crucially, we provide an exact delay distribution analysis characterizing how backlog estimation impacts delay constraints while maximizing throughput.

II. S-ALOHA with Delay Constraint

A. System Model

Suppose an S-ALOHA system, where a base station (BS) is located in the center of the coverage area. Time is slotted

This work was supported in part by Brain Pool program funded by the Ministry of Science and ICT through the National Research Foundation of Korea (RS-2025-25456394) and in part by the Science and Engineering Board under grant CRG/2023/005421. *(Corresponding author: Hu Jin, and Jun-Bae Seo.)*

Z. Cao and H. Jin are with the Department of Electrical and Electronic Engineering, Hanyang University, Ansan 15588, South Korea (e-mail: zycao@hanyang.ac.kr, hjin@hanyang.ac.kr).

S. De is with the Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, 110016, India (e-mail: swadesd@ee.iitd.ac.in).

J.-B. Seo is with the Department of Artificial Intelligence and Information Engineering, Gyeongsang National University, Jinju 53064, Republic of Korea (e-mail: jbseo@gnu.ac.kr).

with equal duration, where each slot corresponds to one packet transmission time. Users having one packet to transmit arrive at the system, i.e., becoming backlogged, according to a Poisson process with the mean rate λ (packets/slot). The users can have only one packet. Thus, users and packets are indistinguishable.

At every slot boundary, the BS broadcasts the feedback message that includes the outcome of the previous slot (collision or success) and a (re)transmission probability p_t. Then, backlogged users draw a random number from the unit interval, and (re)transmit their packet at the each slot boundary if the random number is less than p_t. A collision occurs when more than one users (re)transmit in the same slot. The transmission outcome is again immediately fed back to the backlogged users over a separate downlink channel.

B. Optimization Problem

Before introducing an optimization problem for access delay, let us introduce some notations and their definitions. At the *beginning* of slot t, let $N_t \in \mathbb{Z}_{\geq 0}$ denote the backlog size. Furthermore, $S_{t-1} \in \{0, 1\}$ and $A_{t-1} \in \mathbb{Z}_{\geq 0}$ denote the success indicator (the number of packets successfully (re)transmitted) and the number of new arrivals joining the backlog at slot $t-1$, respectively. At the beginning of slot t, the backlog evolves as

$$N_t \;=\; N_{t-1} \;-\; S_{t-1} \;+\; A_{t-1}, \tag{1}$$

which forms a discrete-time Markov process. If the BS knows that $N_{t-1} = n$ backlogged users compete for the channel, it broadcasts a state-dependent retransmission probability $p_t = \hat{p}_n$, which determines the transmission attempt probability for each user. Let $b_i(n)$ denote the probability that i packets are (re)transmitted when the backlog size is n:

$$b_i(n) = \binom{n}{i}(\hat{p}_n)^i(1-\hat{p}_n)^{n-i}. \tag{2}$$

The policy $\hat{p}_n = 1/n$ maximizes the probability of exactly one successful transmission, i.e., $b_1(n)$. Taking the expectation of (1) and averaging over T slots (using telescoping sum), we get $\frac{1}{T}(\mathbb{E}[N_T] - \mathbb{E}[N_0]) = \frac{1}{T}\sum_{t=1}^{T}\big(\mathbb{E}[A_{t-1}] - \mathbb{E}[S_{t-1}]\big)$. As $T \to \infty$, the left-hand side (LHS) vanishes for an ergodic system, yielding $\mathbb{E}[A_{t-1}] = \mathbb{E}[S_{t-1}]$. This shows the output rate $\lambda_{\text{out}} = \sum_{n\geq 0} b_1(n)\pi_n = \lambda$, where $\pi_n = \lim_{t\to\infty}\Pr[N_t = n]$ is the stationary backlog distribution. By Little's law, the mean access delay is expressed as

$$\overline{D} = \frac{\mathbb{E}[N_\infty]}{\lambda_{\text{out}}} = \frac{\sum_{n\geq 0} n\,\pi_n}{\sum_{n\geq 0} b_1(n)\pi_n} = \frac{\sum_{n\geq 0} n\,\pi_n}{\lambda}. \tag{3}$$

This explicitly shows that delay, backlog size, and throughput are linked through rate conservation.

Proposition 1. *For a fixed input rate λ and retransmission policy $\hat{p}_n$, minimizing the mean access delay $\overline{D}$ is equivalent to minimizing the mean backlog $\mathbb{E}[N_\infty]$.*

Proof: By (3), $\overline{D} = \mathbb{E}[N_\infty]/\lambda$ for fixed λ, so minimizing $\overline{D}$ directly minimizes $\mathbb{E}[N_\infty]$. ■

Proposition 2. *Suppose that an arrival process $\{A_t\}$ and an initial backlog N_0 are fixed. Consider two policies that generate $\{S_t^{(1)}\}$ and $\{S_t^{(2)}\}$ such that $S_t^{(1)} \geq S_t^{(2)}$ for all t. Let the corresponding backlogs evolve by $N_t^{(i)} = N_{t-1}^{(i)} - S_{t-1}^{(i)} + A_{t-1}$, $i \in \{1, 2\}$, driven by the* same *arrival realizations $\{A_t\}$. Then, for all t, $N_t^{(1)} \leq N_t^{(2)}$. Consequently, $\mathbb{E}[N_t^{(1)}] \leq \mathbb{E}[N_t^{(2)}]$ for all t, and if both systems are positive recurrent, $\mathbb{E}[N_\infty^{(1)}] \leq \mathbb{E}[N_\infty^{(2)}]$.*

Proof: For $t = 1$, we have $N_1^{(1)} - N_1^{(2)} = (N_0^{(1)} - N_0^{(2)}) - (S_0^{(1)} - S_0^{(2)}) + (A_0 - A_0) \leq N_0^{(1)} - N_0^{(2)} = 0$ (the same initial backlog N_0). If $N_{t-1}^{(1)} \leq N_{t-1}^{(2)}$, then $N_t^{(1)} - N_t^{(2)} = (N_{t-1}^{(1)} - N_{t-1}^{(2)}) - (S_{t-1}^{(1)} - S_{t-1}^{(2)}) + (A_{t-1} - A_{t-1}) \leq 0$, since $S_{t-1}^{(1)} \geq S_{t-1}^{(2)}$. Thus, $N_t^{(1)} \leq N_t^{(2)}$ for all t. Taking expectations yields the stated inequalities. ■

The optimization problem of guaranteeing the mean access delay is formulated as

$$\begin{aligned} \underset{\lambda}{\text{maximize}} \quad & \sum_{n=0}^{\infty} b_1(n)\pi_n, \\ \text{subject to} \quad & \sum_{n=0}^{\infty} \pi_n = 1, \quad \text{and} \quad \overline{D} \leq d_{\max}. \end{aligned} \tag{4}$$

The objective function of (4) is the throughput $\mathbb{E}[S_\infty]$, whereas the first constraint indicates the stable system, i.e., ergodicity of the Markov process N_t. The second constraint guarantees the mean access delay $\overline{D}$ in (3) less than a threshold $d_{\max}$, which is based on Little's law.

Proposition 3. *For a given retransmission policy $\hat{p}_n$ (e.g., $\hat{p}_n = 1/n$), the optimal λ in* (4) *is achieved when the delay constraint is* active*: $\sum_{n\geq 0}\big(n - d_{\max} b_1(n)\big)\pi_n = 0$, and the optimizer is unique.*

Proof: Propositions 1–2 show that minimizing $\overline{D}$ is equivalent to minimizing $\mathbb{E}[N_t]$, which corresponds to maximizing the mean number of successful (re)transmissions $\mathbb{E}[S_{t-1}]$ via (1). Thus, λ can be increased until $\overline{D} = d_{\max}$, making the delay constraint tight. For uniqueness, define the one-step conditional drift [12] as $\mathcal{D}_n \triangleq \mathbb{E}[N_t - N_{t-1} \mid N_{t-1} = n]$, which is the expected net change in backlog given the current backlog n. For some scalar $\delta > 0$ and integer n^*, the Markov chain admits a stationary distribution if $\mathcal{D}_n \leq -\delta$ for all $n > n^*$. Using (1), we can rewrite $\mathcal{D}_n$ as $\mathcal{D}_n = \mathbb{E}[A_{t-1}|N_{t-1} = n] - \mathbb{E}[S_{t-1}|N_{t-1} = n] = \lambda - b_1(n)$. When $\hat{p}_n = 1/n$, $b_1(n)$ decreases strictly with n, ensuring that both n^* and the corresponding λ are unique. ■

The problem in (4) can not be solved by a conventional optimization problem, because π_n depends on λ and $\hat{p}_n$. To find π_n, we first express the relation between π_0 and π_1 using the balance equation:

$$\pi_0 = a_0\pi_0 + a_0 b_1(1)\pi_1. \tag{5}$$

Due to our assumption on Poisson arrivals, we have $a_k = \frac{\lambda^k}{k!}e^{-\lambda}$. For $n \geq 1$, the balance equation for π_n is

$$\pi_n = a_n\pi_0 + \sum_{k=0}^{n-1} c_k(n)\pi_{n-k} + a_0 b_1(n+1)\pi_{n+1}, \tag{6}$$

where $c_k(n) \triangleq a_{k+1}b_1(n-k) + a_k(1-b_1(n-k))$. To find π_n numerically, let us rewrite (5) as $\pi_1 = \frac{1-a_0}{a_0 b_1(1)}\pi_0$. As π_1 can be expressed in terms of π_0, we can write π_n for $n \geq 2$ with respect to π_0 substituting π_k for $k \leq n$ into (6). Let $\hat{\pi}_n$ be the probability of state n when π_0 is set to one, e.g., $\hat{\pi}_1 = \frac{1-a_0}{a_0 b_1(1)} = e^{\lambda} - 1$, and $\pi_n = \hat{\pi}_n \pi_0$. We can get $\hat{\pi}_n$ iteratively (using (6)) as

$$\hat{\pi}_{n+1} = \frac{1}{a_0 b_1(n+1)}\Big(\hat{\pi}_n - a_n - \sum_{k=0}^{n-1} c_k(n)\hat{\pi}_{n-k}\Big). \tag{7}$$

Using $\sum_{n=0}^{\infty} \pi_n = 1$, we determine $\pi_0 = \left(1 + \sum_{n=1}^{N_{\max}} \hat{\pi}_n\right)^{-1}$ and $\pi_n = \hat{\pi}_n \pi_0$, where we will show how to determine $N_{\max}$ later.

To find the convergence condition of π_n, if $\hat{p}_n = p$ in (2), then $\mathcal{D}_n > 0$ as $n \to \infty$. This means that the system becomes unstable when a retransmission probability independent of the state is used. If $\hat{p}_n = \frac{1}{n}$ (maximizer for $b_1(n)$) and $n \to \infty$, it follows that $\lambda < \left(1 - \frac{1}{n}\right)^{n-1} \approx e^{-1}$. If estimation induces a multiplicative bias $c > 0$ so that $p_n = 1/(cn)$, then $\lim_{n\to\infty} b_1(n) = \frac{1}{c}e^{-1/c}$, yielding a sufficient stability bound $\lambda < \frac{1}{c}e^{-1/c}$. More explicitly, let us examine the ratio of π_{n+1} to π_n: $\frac{\pi_{n+1}}{\pi_n} = \frac{1}{a_0 b_1(n+1)}\left(1 - a_n\frac{\pi_0}{\pi_n} - \sum_{k=0}^{n-1} c_k(n)\frac{\pi_{n-k}}{\pi_n}\right)$. If we impose the condition of $\frac{\pi_{n+1}}{\pi_n} < 1$, we get

$$1 - a_0 b_1(n+1) < a_n \frac{\pi_0}{\pi_n} + \sum_{k=0}^{n-1} c_k(n)\frac{\pi_{n-k}}{\pi_n}. \tag{8}$$

As a sufficient condition for $\frac{\pi_{n+1}}{\pi_n} < 1$, we can focus on the following term from (8): $1 - a_0 b_1(n+1) < a_n \frac{\pi_0}{\pi_n}$, which can be expressed as

$$\frac{\pi_n}{\pi_0} < \frac{a_n}{1 - a_0 b_1(n+1)}. \tag{9}$$

For $\lambda < b_1(n) \approx b_1(n+1) \approx e^{-1}$ for $n \to \infty$, we can see that the ratio in (9) vanishes, since $a_n \to 0$ for $n \to \infty$:

$$\frac{\pi_n}{\pi_0} < \frac{a_n}{1 - e^{-(\lambda+1)}} = \frac{\lambda^n e^{-\lambda}}{n!(1 - e^{-(\lambda+1)})} < \eta, \tag{10}$$

where η helps us to predict that $\hat{\pi}_n = \eta\pi_0$ for $\pi_0 = 1$. Since $\pi_0 < 1$, π_n is smaller than η, i.e., $\pi_{N_{\max}} = \hat{\pi}_{N_{\max}}\pi_0$. This enables us to find $N_{\max}$ that makes π_n negligibly small. Let us assume that $\pi_0 = 1$ to find $\hat{\pi}_n$. Since λ is not greater than e^{-1}, we can see that $\hat{\pi}_n$ rapidly diminishes in (10) as $n \to \infty$.

For $\hat{p}_n = \frac{1}{n}$, we can find the solution of (4) by using a bisection search for $\lambda \in (0, e^{-1})$ as shown in Algorithm 1. In Algorithm 1, λ_l and λ_r are the lower/upper bound values for λ. We initialize $\lambda_l = 10^{-3}$ (small but positive) and $\lambda_r = e^{-1} - 10^{-2}$ because, for S-ALOHA with $\hat{p}_n = 1/n$, the throughput per slot is at most $1/e$ (since $b_1(n) = (1 - \frac{1}{n})^{n-1} \leq e^{-1}$). Any bound $0 < \lambda_l < \lambda^\star < \lambda_r < 1/e$ guarantees bisection convergence.

C. Backlog Estimation

Algorithm 1 Bisection search for the solution

```
Initialization: $\lambda_l = 10^{-3}$ and $\lambda_r = e^{-1} - 10^{-2}$.
Output: the maximum arrival rate $\lambda$
while $|\lambda_r - \lambda_l| > \epsilon_b$ do
    $\lambda_m = 0.5(\lambda_l + \lambda_r)$
    Compute $\pi_n$ using (7) with $a_k = \frac{\lambda_m^k}{k!}e^{-\lambda_m}$
    if $\sum_{n=0}^{N_{\max}} n\pi_n > d_{\max}\sum_{n=0}^{N_{\max}} b_1(n)\pi_n$ then
        $\lambda_r = \lambda_m$
    else
        $\lambda_l = \lambda_m$
```

To guarantee the mean access delay subject to $d_{\max}$ and stabilize the system, it is essential to estimate the number of backlogged users n as accurately as possible. To do this, we develop an EKF backlog estimation algorithm presented in Algorithm 2. The notation is summarized as follows: $\hat{N}_{t|t-1}$ and $\hat{N}_t$ denote the predicted state estimate, and the updated state estimate at time t, after observing $\mathbf{z}_t$. The observation vector $\mathbf{z}_t$ is expressed as $\mathbf{z}_t = \left[Z_{1,t},\, Z_{2,t}\right]^T \in [0,1]^2$, where $Z_{1,t}$ and $Z_{2,t}$ indicate whether the slot outcome at time t is *Idle* or *Success*, respectively, and the superscript T denotes the transpose of a matrix or vector. Additionally, we set $Z_{1,t} + Z_{2,t} \in \{0,1\}$, whereas $\mathbf{z}_t = [1,0]^T$ is idle, $\mathbf{z}_t = [0,1]^T$ success; a collision maps to $\mathbf{z}_t = [0,0]^T$. Furthermore, $P_{t|t-1}$ and P_t denote the predicted and updated covariances of the estimation error at time t, after observing $\mathbf{z}_t$. The $\mathbf{K}_t$ denotes the optimal Kalman gain at time t.

1) Algorithm of providing the mean access delay: To apply an EKF algorithm, let us rewrite (1) as $N_t = N_{t-1} - S_{t-1} + A_{t-1} + w_{t-1}$, where $w_{t-1} \sim \mathcal{N}(0, Q_t)$ is a Gaussian process with zero mean and variance Q_t showing the process noise; Q_t is conservatively set to 1 to represent the magnitude of disturbance, since the mean of packet arrivals at a slot in a stable S-ALOHA system generally does not exceed 0.5.

The observation model of the proposed EKF, i.e., the BS observes $\mathbf{z}_t$, is expressed as

$$\mathbf{z}_t = \mathbf{h}(N_t) + \mathbf{v}_k, \tag{11}$$

where $\mathbf{v}_k \sim \mathcal{N}(\mathbf{0}, R_t)$ is the observation noise with covariance matrix R_t. In (11), let us rewrite $\mathbf{h}(N_t)$ as $\mathbf{h}$ for the simplicity of notations and define $\mathbf{h} = [h_1, h_2]$, where h_1 and h_2 are the probability that the outcome of slot t is idle, and success, respectively, for N_t backlogged users with retransmission probability p_t at slot t. We can write h_1 and h_2 as

$$h_1 = (1 - p_t)^{N_t}, \quad \text{and} \quad h_2 = N_t\, p_t\, (1 - p_t)^{N_t - 1}. \tag{12}$$

Using (12), the measurement-noise covariance R_t is taken as a 2×2 matrix with a small ridge εI_2 (I_2 is an identity matrix of size two) to ensure the positive definiteness [13]:

$$R_t = \begin{bmatrix} h_1(1-h_1) & -h_1 h_2 \\ -h_1 h_2 & h_2(1-h_2) \end{bmatrix} + \varepsilon I_2. \tag{13}$$

The Jacobian of $\mathbf{h}$ is $\mathbf{H}_t = \left[\partial h_1/\partial N_t\, \partial h_2/\partial N_t\right]^T$, which is a linearization of the non-linear $\mathbf{h}$ in (11) for running the Kalman filter. Differentiating (12) with respect to N_t yields:

$$\begin{aligned} \frac{\partial h_1}{\partial N_t} &= h_1 \ln(1 - p_t), \quad \text{and} \\ \frac{\partial h_2}{\partial N_t} &= p_t(1 - p_t)^{N_t - 1}\big[1 + N_t \ln(1 - p_t)\big]. \end{aligned} \tag{14}$$

Algorithm 2 EKF-Based Backlog Estimation for S-ALOHA

```
1: Initialization: $Q = 1$, $\hat{N}_0 = 1$, $P_0 = 1$, $\varepsilon = 10^{-4}$, $p_t = 1$
2: for each slot $t = 1, 2, \dots$ do
3:     Compute prediction:
       $\hat{N}_{t|t-1} = \hat{N}_{t-1} - S_{t-1}, \quad P_{t|t-1} = P_{t-1} + Q_t$
4:     Compute predicted measurement:
       $\widehat{\mathbf{h}}_t = \mathbf{h}\big(\hat{N}_{t|t-1}\big)$ using (12)
5:     Compute Jacobian $\mathbf{H}_t$ by (14)
6:     Compute Kalman gain:
       $\mathbf{S}_t = \mathbf{H}_t\, P_{t|t-1}\, \mathbf{H}_t^T + R_t, \quad \mathbf{K}_t = P_{t|t-1}\, \mathbf{H}_t^T\, \mathbf{S}_t^{-1}$
7:     Update state and covariance:
       $\hat{N}_t = \hat{N}_{t|t-1} + \mathbf{K}_t\big(\mathbf{z}_t - \widehat{\mathbf{h}}_t\big)$, $P_t = \big(1 - \mathbf{K}_t\mathbf{H}_t\big)P_{t|t-1}$
8:     Broadcast ACB factor $p_t = \min\{1,\ 1/\hat{N}_t\}$
```

To avoid singularities, we limit p_t at $1 - \varepsilon$ when computing $\mathbf{H}_t \in \mathbb{R}^{2\times 1}$.

2) EKF for online access control: In Algorithm 2, we first initialize the process noise covariance $Q_t = 1$, the initial backlog estimate $\hat{N}_t = 1$, the updated covariance $P_t = 1$ at $t = 0$, $\varepsilon = 10^{-4}$, and retransmission probability $p_t = 1$. At the beginning of each slot t, we predict the backlog by subtracting the observed successes S_{t-1} from the previous estimate and update the variance by adding Q_t. The lines 4-7 are the computational procedure of a standard EKF. After estimating $\hat{N}_t$, the BS broadcasts $p_t = \min\{1, 1/\hat{N}_t\}$.

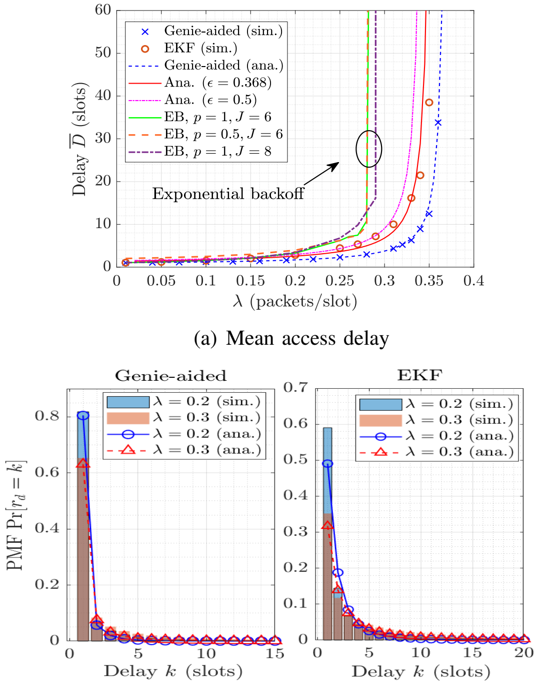
3) Computational complexity: In each slot t, the BS executes Algorithm 2, which includes prediction, Jacobian computation, Kalman gain and covariance update. These steps impose a fixed computation load independent of the backlog size; about $12 - 16$ scalar operations in total, and at most *two* 2×2 matrix operations (one small multiply/accumulate for $\mathbf{S}_t, \mathbf{K}_t$ and one 2×2 inversion). Hence, the EKF incurs constant time and memory complexity $O(1)$.

D. Access Delay Distribution

To analyze the access delay when $\hat{p}_n = 1/n$, we use a *tagged* user. According to Poisson Arrivals See Time Average (PASTA) property, π_n is the probability that the tagged user sees n backlogged users upon its arrival. In our system, the tagged user will start to transmit its packet in the next slot, which is known as *delayed first transmission*. Let i be the number of other backlogged users at arrival of the tagged user. As mentioned earlier, the tagged user finds i backlogged users with probability π_i, say slot $t - 1$. In that slot, where the tagged user does not transmit, the number of backlogged users becomes n one slot later as $n = (i - S_{t-1}^{\text{others}}(i) + A_{t-1}) + 1$, where $S_{t-1}^{\text{others}}(i) \in \{0, 1\}$ denotes a successful transmission of a packet among i backlogged users; that is, $S_{t-1}^{\text{others}}(i) = b_1(i)$.

Let r_d denote the random variable of representing the access delay of the tagged user. To find $\Pr[r_d = k]$, we consider an absorbing Markov chain, where the absorption state represents the successful transmission of the tagged user's packet. Let ϕ_n be the initial probability that the system has n backlogged users, including the tagged user. We can get ϕ_n as

$$\phi_n = \sum_{i\geq 0} \pi_i \left[(1 - b_1(i))a_{n-1-i} + b_1(i)a_{n-i}\right]. \tag{15}$$



(a) Mean access delay

(b) PMF of access delay ($\epsilon = 0.5$ for EKF)

Fig. 1: Delay performance.

For example, ϕ_1 shows the probability that the tagged user finds the system empty (with probability π_0) and no other arrivals with probability a_0. In other words, ϕ_1 is the probability of the tagged user alone in the system. The initial state probability vector $\vec{\phi} = [\phi_n]$ for $n \geq 1$ shows the state that the tagged user is with $n-1$ backlogged users in the system. The access delay is the time for the tagged user to move to the absorption state, i.e., absorption time. The probability mass function (PMF) of RA delay for k slots can be expressed as

$$\Pr[r_d = k] = \vec{\phi} \cdot H^{k-1} U. \tag{16}$$

The elements of the (sub)matrix U indicate the state transition probabilities to the absorbing state (state 0) from other states:

$$U^T = \begin{array}{cc} \text{state} & \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & \cdots \end{array} \\ 0 & \left[\begin{array}{cccccc} 1 & 1 & \gamma_2 & \gamma_3 & \gamma_4 & \cdots \end{array}\right] \end{array}. \tag{17}$$

where $\gamma_n = \hat{p}_n\,(1 - \hat{p}_n)^{n-1} = \frac{1}{n}\left(1 - \frac{1}{n}\right)^{n-1}$. In (17), at state 1, where the tagged user alone is in the system, it can make a successful transmission with probability 1. This γ_n shows the probability that the tagged user (re)transmit with probability $\hat{p}_n$, while the other $n - 1$ users will not. In (16), we write matrix $H = [\alpha_{n,k}]$ for $n, k \geq 1$, where each element $\alpha_{n,k}$ indicates the state transition probability in the absorbing Markov chain. We have $\alpha_{1,k} = 0$ for $k \geq 1$ and $\alpha_{n,k} = 0$ for $k < n - 1$. Moreover, $\alpha_{n,n-1}$ for $n \geq 2$ is expressed as $\alpha_{n,n-1} = (1 - \hat{p}_n)\,\beta_n a_0 = \left(1 - \frac{1}{n}\right)\beta_n a_0$, and $\beta_n = \binom{n-1}{1}\hat{p}_n\,(1 - \hat{p}_n)^{n-2} = \binom{n-1}{1}\frac{1}{n}\left(1 - \frac{1}{n}\right)^{n-2}$. For $k \geq n$, we can write $\alpha_{n,k}$ as $\alpha_{n,k} = (1 - 1/n)\,(\beta_n a_{k-n+1} + (1 - \beta_n)a_{k-n})$. The mean access delay $\overline{D}$ can be obtained by either the second constraint of (4), or

$\overline{D} = \vec{\phi}(I - H)^{-1}\mathbf{1}$ (using (16)), where a proper truncation for matrix H is needed and $\mathbf{1}$ is a column vector of ones.

III. Numerical Results

While the optimization problem in (4) assumes perfect knowledge of backlog size, the EKF produces estimation errors. Consequently, even when we constrain the input rate λ to the solution of (4), the access delay limit cannot be satisfied. Therefore, we need to tune p_n in a practical manner that accounts for estimation errors. We set $N_{\max} \sim 600$, $\epsilon_b = 10^{-4}$ in Algorithm 1, and truncate the size of matrix H by 800. In [11], an exponential backoff (EB) algorithm is proposed to reduce age-of-information (AoI). A backlogged user transmits a packet with probability $p\alpha^{i-1}$ for $i \in \{1, \dots, J+1\}$ for the ith attempt. For $i = J + 1$, the user repeatedly uses $p\alpha^J$. For a given pair of p and J, α is optimized to minimize AoI in [11]. The AoI just before a new update arrives (i.e., the peak AoI) reflects the combined waiting and access delays of both the previous and the upcoming packets. To apply it to our system, we numerically find the optimal α that minimizes the access delay of a packet for each (λ, p, J) through exhaustive searching. To address this, Fig. 1(a) compares the mean access delay $\overline{D}$ of the EKF-based backoff algorithm and the Genie-aided system. The Genie-aided system has perfect knowledge of the exact backlog size at each slot, enabling it to implement $r_n = 1/n$ without error. Since our analysis uses r_n (based on the exact backlog size), the analytical results agree well with simulation results of the Genie-aided system for both the mean access delay in Fig. 1(a) and PMF in Fig. 1(b).

In order to reflect the estimation errors by EKF for the optimization problem, we modify $\hat{p}_n$ in (4) as $\hat{p}_n \simeq \frac{1}{n+\epsilon n} = \frac{1}{cn} \simeq p_t$, which is called a scaled $\hat{p}_n$. Notice that ϵ (or $c = 1 + \epsilon$) is a random variable that depends on various factors such as each state and λ. It can take positive or negative values. When ϵ takes a negative value, which means that EKF underestimates the backlog size, more collisions can occur. A positive ϵ means an overestimation on backlog size, implying more delay. The key insight of the tuned $\hat{p}_n$ is that function $b_1(n) = n\hat{p}_n(1 - \hat{p}_n)^{n-1}$ takes almost symmetric shape around $\hat{p}_n = 1/n$; that is, for small ϵ, $\frac{n}{n-\epsilon n}(1 - \frac{1}{n-\epsilon n})^{n-1} \approx \frac{n}{n+\epsilon n}(1 - \frac{1}{n+\epsilon n})^{n-1}$. Thus, we simplify it using a positive constant c for practical use. Fig. 1(a) shows the analytical results with two values of ϵ: 0.368 and 0.5. While the mean access delay with $\epsilon = 0.368$ shows a good fit with simulation, $\epsilon = 0.5$ conservatively works well. This is also observed in Fig. 1(b); the PMF of the EKF-based backoff algorithm with $\epsilon = 0.5$ shows slightly longer delays against simulation. It is notable that we have $\lim_{n\to\infty} b_1(n) = \lim_{n\to\infty} n\frac{1}{cn}\left(1 - \frac{1}{cn}\right)^{n-1} = \frac{1}{c}e^{-1/c}$. Compared to the EB algorithm in [11], Fig. 1(a) shows that our proposed algorithm achieves consistently lower delay. The performance gap between two algorithms becomes large as λ increases, particularly in the high-load case; the proposed algorithm outperforms the EB in delay.

In Table I, we present the solution of (4), denoted by λ^*, for given delay limits $d_{\max}$. It is obtained using Algorithm 1 and $\hat{p}_n = 1/(1.5n)$. Then, λ^* is used as an input to the system with the EKF-based backoff algorithm. We then get $\overline{D}$ through simulation for a given λ^* to see that $\overline{D}$ is kept below $d_{\max}$. For various values of $d_{\max}$, we can see that $\overline{D}$ is conservatively maintained close to $d_{\max}$.

TABLE I: Measured mean access delay vs. delay constraint

$d_{\max}$	2.5	3.5	4.5	5.5	6.5	7.5	8.5
λ^*	0.168	0.225	0.253	0.271	0.283	0.291	0.297
$\overline{N}$ (ana.)	0.486	0.859	1.220	1.578	1.931	2.286	2.525
$\overline{N}$ (sim.)	0.377	0.758	1.170	1.477	1.840	2.115	2.416
$\overline{D}$ (sim.)	2.243	3.372	4.614	5.472	6.506	7.271	8.124

IV. Conclusions

This work investigated an optimization framework that guaranteed the mean access delay of S-ALOHA for a given delay limit. The insight of the optimization is the equivalence between throughput maximization and delay minimization, providing a unified approach to S-ALOHA optimization. To find the solution, the S-ALOHA system was extensively analyzed, including the PMF of the access delay. The principle of guaranteeing the mean access delay is to maximize the throughput while allowing input rate up to the input rate limit, where the information on the exact backlog size for every slot is utilized. The limit is that, since the estimation errors on the backlog size are unavoidable, the (re)transmission probability is empirically tuned to meet the mean delay limit.

References

[1] S. C. Liew, Y. J. Zhang, and D. R. Chen, "Bounded-mean-delay throughput and nonstarvation conditions in Aloha network," *IEEE/ACM Transactions on Networking*, vol. 17, no. 5, pp. 1606-1618, Oct. 2009.

[2] J.-B. Seo, and V. C. M. Leung, "Queuing performance of multichannel S-ALOHA systems with correlated arrivals," *IEEE Trans. Veh. Technol.*, vol. 60, no. 9, pp. 4575-4586, Nov. 2011.

[3] Y. Li, W. Zhan, and L. Dai, "Rate-constrained delay optimization for Slotted Aloha," *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5283-5298, Aug. 2021.

[4] F. A. Tobagi, "Distributions of packet delay and interdeparture time in slotted ALOHA and carrier sense multiple access," *J. ACM*, vol. 29, no. 4, pp. 907–927, Oct. 1982.

[5] Y. Yang, and T.-S.P. Yum, "Delay distributions of slotted ALOHA and CSMA," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1846–1857, Nov. 2003.

[6] B. J. Kwak, N.-O. Song, and L. E. Miller, "Performance analysis of exponential backoff," *IEEE/ACM Trans. Netw.*, vol. 13, no. 2, pp. 343–355, Apr. 2005.

[7] L. Barletta, F. Borgonovo, and I. Filippini, "The throughput and access delay of slotted-aloha with exponential backoff," *IEEE/ACM Trans. Netw.*, vol. 26, no. 1, pp. 451–464, Feb. 2018.

[8] M. E. Rivero-Angeles, D. Lara-Rodriguez, and F. A. Cruz-Perez, "Gaussian approximations for the probability mass function of the access delay for different backoff policies in S-ALOHA," *IEEE Commun. Lett.*, vol. 10, no. 10, pp. 731–733, Oct. 2006.

[9] J.-B. Seo, W. T. Toor, and H. Jin, "Analysis of two-step random access procedure for cellular ultra-reliable low latency communications," *IEEE Access*, vol. 9, pp. 5972-5985, 2021.

[10] X. Sun, W. Zhan, W. Liu, Y. Li and Q. Liu, "Sum rate and access delay optimization of short-packet Aloha," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 1501-1514, 2022.

[11] P. Mollahosseini, S. Asvadi, and F. Ashtiani, "Effect of variable backoff algorithms on age of information in Slotted ALOHA networks," *IEEE Trans. on Mobile Comput.*, vol. 23, no. 9, pp. 8620–8633, Sept. 2024.

[12] D. P. Bertsekas, and R. G. Gallager, *Data Networks*, 2nd Ed., Prentice-Hall, 1992

[13] M. S. Grewal and A. P. Andrews, *Kalman Filtering: Theory and Practice with MATLAB*, 4th Ed., Hoboken, NJ, USA: Wiley-IEEE Press, 2014.