Boosting Slotted Aloha with Successive Transmission: Modeling and Performance Optimization

Weilong Zhu, Wen Zhan, *Member, IEEE*, Xinghua Sun, *Member, IEEE*, Xiang Chen and Yuan Jiang

***Abstract*—How to effectively support massive access and data transmission in Internet of Things scenarios has been a long-standing and critical issue for various wireless communication networks. To address this issue, a flexible and efficient medium access control protocol is the key. In this paper, we propose Slotted Aloha with Successive Transmission (SAST) scheme, in which upon the successful transmission of the Head-of-Line (HoL) packet, the node delivers the remaining packets with probability 1 until the buffer is cleared or a collision occurs, thereby capitalizing on immediate channel availability. By formulating vacation queuing models of both node and channel, the access/data throughput and access/data delay are explicitly characterized and optimized by properly choosing the transmission probability of the HoL packet. Our analysis reveals that the maximum data throughput of SAST scheme is 0.5, higher than e^{-1} in classic slotted Aloha. The practical insights of the analysis are also demonstrated by taking the example of 2-step Small Data Transmission (SDT) random access in 5G. It is shown that the SAST scheme can be seamlessly implemented into 5G and the comparison with 2-step SDT random access reveals that SAST can improve the throughput performance while significantly reduce the signaling overhead, nearly halved in the saturated case and up to 70% reduction in the unsaturated case.**

***Index Terms*—Aloha, random access, successive transmission, vacation queuing model, 5G**

I. Introduction

The prevalence of the Internet of Things (IoT) has profoundly impacted various application domains, including unmanned aerial vehicles (UAVs), industrial wireless sensor networks (IWSNs), and environmental monitoring [1]. In these applications, small data packets, typically only a few hundred bits in size, are the primary form of traffic generated by IoT devices. However, with the continuous evolution and progress of technology and industry, the characteristics of traffic become increasingly complex, and the transmission of essential long data packets in scenarios like smart factories must be given attention [2]. Therefore, a key challenge faced by wireless communication systems is to ensure compatibility with the simultaneous access of a massive number of diverse IoT devices, while also accommodating the transmission of long data packets.

To support diverse traffic characteristics, random access schemes have demonstrated effectiveness and applicability owing to their simplicity and adaptability. These schemes have been widely adopted in cellular systems, WiFi, LoRa, and others. With random access, devices independently and distributively decide when to access the channel. Although various random access schemes have been proposed, they can be broadly categorized into two types based on whether a connection is established beforehand: connection-based and packet-based random access scheme.

In connection-based random access scheme, the device first sends a data transmission request (typically much smaller than a data packet). Only after receiving an acknowledgment from the receiver does it proceed to the data transmission process in a collision-free manner. The connection-based scheme is suitable for scenarios where packets arrive frequently and the length of data packets is much larger than that of the request. Since collisions occur during the request transmission process, the overhead of transmission failures can be significantly reduced. Examples of connection-based random access scheme include LTE [3] and the Request to Send/Clear to Send (RTS/CTS) mechanism in WiFi [4]. However, in the context of IoT communication, small data transmission has become mainstream, rendering connection-based random access scheme inefficient due to excessive connection establishment overhead. In such cases, the packet-based scheme is a more appropriate solution. In contrast to connection-based scheme, the packet-based scheme allows devices to directly transmit its data packet without connection establishment, where the overhead is entirely determined by the length of data packet. To support Small Data Transmissions (SDT), 3GPP incorporated the data transmission into the random access procedure and presented

Manuscript received May 25, 2024; revised Oct. 16, 2024 and Feb. 1, 2025; accepted Feb 15, 2025. The work of X. Chen was supported by National Key R&D Program of China 2022YFB2902002. The work of X. Sun was supported by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515012015. The work of W. Zhan was supported in part by The Shenzhen Science and Technology Program (No.RCBS20210706092408010), in part by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University, under Grant 24qnpy204, in part by Open Fund of State Key Laboratory of Satellite Navigation System and Equipment Technology (No. CEPNT-2021KF-04) and in part by the Science and Technology Planning Project of Key Laboratory of Advanced IntelliSense Technology, Guangdong Science and Technology Department under Grant 2023B1212060024. (*Corresponding authors: Wen Zhan.*)

This paper was presented in part at The International Conference on Wireless Communications and Signal Processing, Hefei, Anhui, China, October 2024.

Weilong Zhu, Wen Zhan, Xinghua Sun and Yuan Jiang are with the School of Electronics and Communication Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China (e-mail:zhuwlong5@mail2.sysu.edu.cn;zhanw6@mail.sysu.edu.cn; xsunxinghua@mail.sysu.edu.cn;jiangyuan3@mail.sysu.edu.cn).

Xiang Chen is with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, Guangdong 510006, China (e-mail: chenxiang@mail.sysu.edu.cn).

the 2/4-step SDT schemes in Release 17 [5], which is effective in reducing signaling overhead and energy consumption [6], [7]. While the aforementioned developments have been substantial, the packet-based random access schemes often hit the performance bottleneck in the presence of massive access requests or long packets, where the throughput becomes low and network congestion is inevitable. The fundamental reason lies in the insufficient applicability of current random access mechanisms when faced with the complicated data traffic characteristics of IoT scenarios. In this paper, a new and scalable random access scheme named Slotted Aloha with Successive Transmission (SAST) is proposed. In the SAST scheme, the initial transmission step of the Head-of-Line (HoL) packet is the same as in the 4/2-step SDT RA scheme. Once the HoL packet is successfully transmitted, the remaining packets in the node's buffer are transmitted successively with probability 1 until the buffer queue is cleared or a collision occurs. Our analysis reveals SAST scheme can achieve the maximum data throughput of 0.5, higher than that of classic slotted Aloha, while the SAST scheme is compatible with the SDT scheme proposed by 3GPP R17.

To investigate the performance of the SAST scheme, we characterize the behavior of node and channel by leveraging the discrete-time vacation queuing theory. From the perspective of node, successive transmission indicates a busy period in the queueing system, while the vacation period of node indicates its buffer is either empty, or non-empty but it does not transmit, or a collision occurs. From the perspective of channel, the busy period is the time period in which packets are transmitted successfully; otherwise, it is in the vacation period. By incorporating the vacation queuing analysis of both node and channel, we derive the mean length of busy/vacation period given the system input parameters including the packet arrival rate and the transmission probability of HoL packet, enabling the further analysis of throughput, delay and the signalling overhead of the SAST scheme. The main contributions of the paper are summarized below:

- We propose Slotted Aloha with Successive Transmission (SAST). The main idea of SAST is to enable successive transmission by nodes upon the HoL packet is successfully transmitted. Since no change is needed at the PHY layer and the signaling exchange process at the MAC/RRC layer, the SAST scheme is compatible with the existing 4/2-step SDT scheme.
- We establish two vacation queuing models to analyze the throughput performance of SAST from the perspective of channel and that of node, respectively. The access delay, packet delay, access throughput and data throughput are explicitly derived as functions of key system parameters including the number of nodes, the input rate and the transmission probability of HoL packet. The maximum access throughput and data throughput and corresponding optimal transmission probabilities are obtained, revealing that the maximum achievable data throughput of SAST scheme is 0.5, obtaining 37% performance gain compared to classic slotted Aloha. This performance gain, as aforementioned, is acquired without any modification to the PHY layer.
- We explain how the proposed SAST can be used in 5G system based on 3GPP MAC specifications from the perspective of signaling exchange. By leveraging the signaling-to-throughput Ratio (STR), i.e., the signaling overhead per successful data packet per slot, we compare the performance of SAST with 2-step SDT. The result shows that the STR of SAST scheme is smaller than that of 2-step SDT scheme, nearly half at most in the saturated case and 70% at most in the unsaturated case.

The remaining sections of the paper are organized as follows. Section II provides an overview of related work in the field. Section III presents a detailed description of the SAST scheme. Section IV formulates vacation queuing models for SAST from perspectives of both node and channel, based on which detailed queueing analysis is presented in Section V. Section VI optimizes the access throughput and data throughput. Section VII employs a two-dimensional Markov chain model to analyze the access delay and packet delay. Section VIII demonstrates the application of the SAST scheme in practical 5G scenarios and compares it with the 2-step SDT scheme in terms of signaling-to-throughput ratio. The extension of the proposed analytical framework to incorporate practical limitations is discussed in Section IX. Finally, Section X concludes the paper by summarizing the key findings.

II. Related Work

The throughput performance limits of classic slotted Aloha with channel collision model[1] has long been known to be e^{-1} [8]–[11]. Yet, how to achieve the maximum throughput depends on parameter settings and the network scenarios.

The first solution is traffic scheduling, which prompts nodes with packet collisions to defer their access requests for a random period before retransmission, thereby smoothing the temporal distribution of channel traffic, reducing concurrent access requests within each time slot, and therefore alleviating transient congestion. Traffic scheduling is usually accomplished via backoff schemes, such as the Geometric Backoff (GB) scheme [12]–[14], the Exponential Backoff (EB) scheme [12]–[16], and other backoff schemes [12], [13]. In [8], [9], [12], by assuming the number of requests in the channel follows Poisson distribution with parameter G, how various backoff algorithms affect the throughput has been analyzed. For bursty arrival scenarios, [17]–[19] adopted the Beta distribution as suggested by 3GPP [20], and analyzed the random access procedure using different backoff schemes in Machine-to-Machine communications by iteratively calculating the number of attempts in each slot.

Besides traffic scheduling, the second solution is resource scheduling, which involves the dynamic allocation of time-frequency/code resources. For instance, in multichannel mechanisms [21]–[24], nodes with packet collision retransmit immediately on a randomly selected channels, thereby collision probability can be reduced as the number of channels increases. In reservation Aloha [25]–[29], nodes first send a

[1]With channel collision model, one packet can be successfully decoded if there is no current transmission.

short reservation packets, and then transmit data if the reservation is successful. Centralized resource scheduling strategy is adopted to ensure that the data transmission process is collision-free. Although traffic/resource scheduling improves the performance of the random access networks in the massive access scenario, the throughput bottleneck remains e^{-1}.

To break the performance bottleneck of classic slotted Aloha, another solution is widely used: advanced receiver structure, which gives the devices the ability of multipacket reception (MPR), i.e., two or more simultaneously transmitted packets can be successfully received. In [30]–[34], it was shown that the throughput in slotted Aloha with MPR could be much larger than e^{-1}. To achieve MPR, [35] proposed an enhanced scheme named Contention Resolution Diversity Slotted Aloha (CRDSA) using successive interference cancellation, which can achieve the maximal throughput approximately to 0.55. Based on CRDSA, more variants [36]–[41] have been proposed. On the other hand, Non-orthogonal multiple access (NOMA) technology, which exploits power domain, code domain (such as code division multiple access), and interleaver pattern for multiple access by assigning different levels to different users over the same time/frequency resources and employing successive interference cancellation at the receiver to mitigate multi-user interference, has garnered widespread attention due to its potential for obtaining better performance than orthogonal multiple access (OMA) [42]–[44]. As demonstrated in [43], the average throughput of NOMA-multichannel Aloha (NM-Aloha) is 1.5 times higher than that of standard multichannel Aloha when only two packets can be simultaneously decoded. Although significant performance can be observed with MPR, the price is a disruptive change of the PHY layer, which could be unpractical for low-cost IoT applications.

In summary, how to efficiently support massive access and data transmission in IoT scenarios has been a long standing issue. Although existing works have gained substantial achievement, a flexible random access scheme, which can break the throughput bottleneck of the classic slotted Aloha while easy to be implemented, still awaits exploration.

III. System Model

Consider a slotted Aloha network containing n nodes and one receiver. The arrival of data packets at each node follows a Bernoulli process with parameter λ and the buffer of each node is infinite[2]. The packets generated by nodes are transmitted over a noiseless channel. All nodes can access the channel only at the beginning of a time slot. We assume perfect and instant feedback from the receiver, such that nodes can be aware of whether their access requests are successful or not by the end of the time slot. The collision model is considered, i.e., each packet can be transmitted successfully only when there is no concurrent transmission.

[2]The analysis in this paper can serve as a good approximation for the SAST network with a large retransmission limit or buffer size. To evaluate the impact of a tight retransmission limit or buffer size on the performance of SAST, the proposed analytical framework should be extended, where the key to this extension lies in characterizing the variation of queue length through an expansion of the Markov chain in Section VII-B.

TABLE I
SUMMARY OF IMPORTANT NOTATIONS USED

Notation	Definition
$\lambda, \hat{\lambda}$	Node input rate, aggregate input rate
q	Transmission probability
θ	Attempt rate
θ_{sa}	Attempt rate in saturated network
B, V, C	Busy period, vacation period, and cycle of node
B_c, V_c, C_c	Busy period, vacation period, and cycle of channel
V_1, V_0	Type-1 vacation period, Type-0 vacation period
p_{ne}	Probability that the buffer of node is non-empty
p_0	Probability that a node can clear its buffer within one busy period
p_A	Probability of successful transmission of access requests
A_B	Number of packet arrivals during B
A_V	Number of packet arrivals during V
A_{V1}, A_{V0}	Number of packet arrivals during V_1 and that during V_0
Q, P	Steady-state queue length at the beginning of B and that at the end of B
L	Steady-state queue length of node buffer

Let us elaborate the SAST scheme. Specifically, the backlogged node transmits the Head-of-Line (HoL) packet in its buffer with probability $q \in (0, 1]$. If the receiver successfully decodes the packet, it will reply with an acknowledgment (ACK) message indicating successful access. Otherwise, a negative acknowledgment (NACK) message is sent to indicate a failed transmission. Here, it is assumed that the ACK/NACK transmission is instantaneous and collision-free. With the successful transmission of the HoL packet, the node will deliver the remaining packets in the following slots with probability 1 until the buffer is cleared or a collision occurs. For the receiver, it replies with an ACK slot by slot, or a NACK when a collision occurs. The HoL packet can be regarded as the initial channel access request. Therefore, the HoL packet and access request are used interchangeably in the following.

In this paper, we evaluate the network performance via the following metrics:

- Access throughput λ_{out}^a: the long-term average number of successful access requests (i.e., the HoL packet in the buffer) per time slot.
- Data throughput λ_{out}^d: the long-term average number of successful packets transmitted per time slot.
- Access delay D_A: the long-term mean time length of access requests from generation to acceptance.
- Packet delay D_P: the long-term mean time length of packets from generation to acceptance.
- Signaling-to-throughput ratio: signaling overhead per successful data packet per slot.

For convenience of description, main mathematical notations are summarized in Table I.

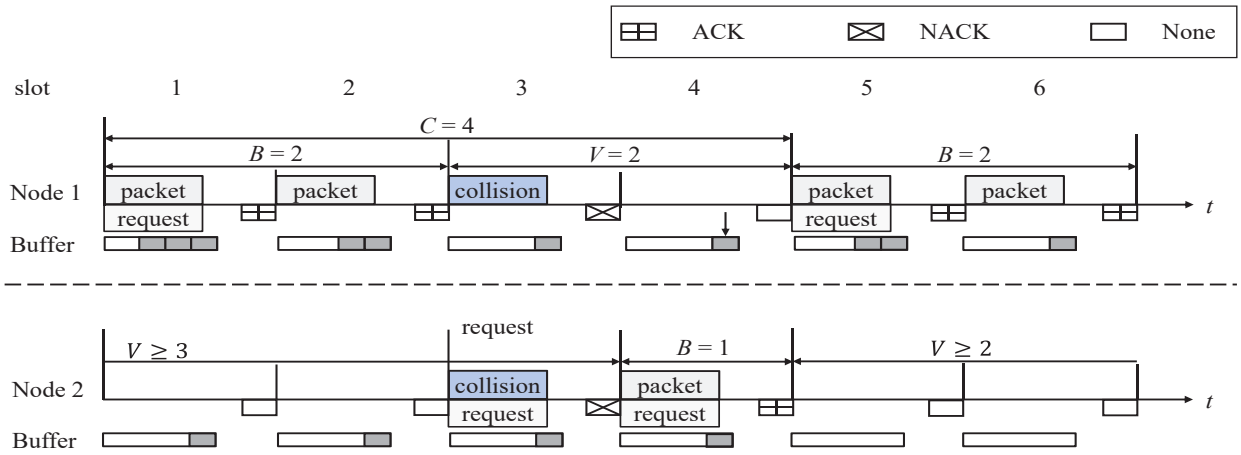

Fig. 1. Illustration of a two-node network with SAST scheme, where B, V, and C denote the busy period, vacation period, and cycle time of node, respectively.

IV. Vacation Queuing Model for SAST

In this section, we formulate vacation queuing model for SAST scheme from the perspective of node and that of channel, respectively.

From the perspective of node, for illustration, Fig. 1 presents the contention process of a two-node case with SAST scheme. Take Node 1 as an example. Node 1 sends its HoL packet at slot 1 (Node 1 has three packets in the buffer) with probability q. With the successful transmission, Node 1 delivers the remaining packets with probability 1 until a collision occurs at slot 3. Because of a new arrival at slot 4, Node 1 has two packets in its buffer and attempt to transmit at the beginning of slot 5. Node 1 clears its buffer at slot 6. Intuitively, the behavior of node can be divided into two alternating periods: busy period and vacation period. Denote B_i as the busy period of node i, during which it transmits its packets successfully, where $i \in \{1, 2, \ldots, n\}$. Denote V_i as the vacation period of node i, during which its buffer is either empty, or non-empty but it does not transmit, or a collision occurs. Denote C_i as a cycle of node i, which is the duration between two consecutive time points that node i sends a batch of packets, where $C_i = B_i + V_i$. As a homogeneous scenario is considered, we drop the subscript i for simplicity.

From the perspective of channel, as illustration in Fig. 2, the busy period and vacation period of channel can also be defined. The busy period of channel, denoted by B_c, is the time period in which packets are transmitted successfully. Both the channel busy period B_c and the node busy period B describe the transmission process of one node, which results in the same probability mass function. Thus, we use the symbol B to represent transmission process. The vacation period of channel, denoted by V_c, is the time period in which no packet is transmitted successfully, that is, either the channel is idle or a collision occurs. Accordingly, we define a cycle of the channel, denoted by C_c, as the duration between two consecutive time points that a batch of packets are beginning to be sent over channel, where $C_c = B_c + V_c$.

Let $\overline{X}$ denote the mean value of the random variable X. Note that in a channel cycle C_c, only one access request can be accepted successfully, i.e., $1/\overline{C}_c$ is the frequency of the successful access requests. On the other hand, the packets can be transmitted successfully only in busy period. Therefore, the proportion of time occupied by the busy period in a channel cycle is the frequency of successful transmissions of packets. According to the definition of throughput, the access throughput and data throughput can be obtained as

$$\lambda_{out}^{a} = \frac{1}{\overline{C}_c} = \frac{1}{\overline{B} + \overline{V}_c} \tag{1}$$

and

$$\lambda_{out}^{d} = \frac{\overline{B}}{\overline{C}_c} = \frac{\overline{B}}{\overline{B} + \overline{V}_c}. \tag{2}$$

In the following section, we derive the mean length of busy period $\overline{B}$ and the mean length of channel vacation period $\overline{V}_c$.

V. Vacation Queuing Analysis for SAST

A. Channel Vacation Queuing Analysis

Let us start the queuing analysis from the perspective of channel. With a large number of nodes n, the number of access requests at each slot, including the newly requests and retransmitted ones, can be approximately regarded as a Poisson random variable[3] with parameter θ, where

$$\theta = nqp_{ne}, \tag{3}$$

where p_{ne} denotes the probability that the buffer of a node is non-empty. θ is also called the attempt rate. Let us first derive the mean length of channel vacation period $\overline{V}_c$ and then the mean length of busy period $\overline{B}$ based on attempt rate θ.

1) Mean length of channel vacation period $\overline{V}_c$: Channel shifts from the busy period B to the vacation period V_c in two cases:

- Case 1: The tagged node clears its buffer and channel shifts to vacation period.

With the attempt rate θ, the channel enters a busy period only when one request is transmitted with probability $\theta e^{-\theta}$. Specifically, upon a tagged node clears its buffer, the length of channel vacation period $V_c = 0$ if another node delivers

[3]The effectiveness of Poisson approximation has been widely verified in the massive random access scenarios [45]–[47].

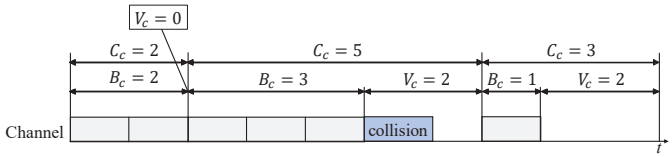

Fig. 2. Illustration of the work process of aggregate channel, where B_c, V_c, and C_c denote the busy period, vacation period, and cycle time of channel, respectively.

its packet successfully in the consecutive slot with probability $\theta e^{-\theta}$, i.e., the channel enters the next busy period immediately. Otherwise, the channel enters a vacation period. By analogy, if $V_c = i$, the channel is either idle or encounters a collision in the previous i slots, and then the channel is occupied successfully by only one node at the $(i+1)$-th slot with probability $(1-\theta e^{-\theta})^i \theta e^{-\theta}$, i.e.,

$$\Pr\{V_c = i|\text{Case 1}\} = \left(1-\theta e^{-\theta}\right)^i \theta e^{-\theta}, i = 0, 1, \dots \quad (4)$$

- Case 2: A collision occurs at the end of busy period and channel shifts to vacation period.

Due to collision, the length of channel vacation period is at least one. Similar to Case 1, the channel vacation period V_c in Case 2 follows a geometric distribution with parameter $\theta e^{-\theta}$ as well. We have

$$\Pr\{V_c = i|\text{Case 2}\} = \left(1-\theta e^{-\theta}\right)^{i-1} \theta e^{-\theta}, i = 1, 2, \dots \quad (5)$$

Denote p_0 as the probability that a transmission node can clear its buffer within one busy period, i.e., the probability of Case 1 is p_0. By combining (4) and (5), we give the mean length of channel vacation period as

$$\begin{aligned}\overline{V}_c &= p_0 E\left[V_c|\text{Case 1}\right] + (1-p_0) E\left[V_c|\text{Case 2}\right] \\ &= p_0 \frac{1-\theta e^{-\theta}}{\theta e^{-\theta}} + (1-p_0)\frac{1}{\theta e^{-\theta}} = \frac{e^{\theta}}{\theta} - p_0.\end{aligned} \quad (6)$$

2) Mean length of channel busy period $\overline{B}$: Let us now derive the mean length of busy period $\overline{B}$ by considering the saturated and unsaturated cases, respectively. [4]

For one hand, in the saturated case, each node always has packets to send, where $p_0 = 0$, $p_{ne} = 1$. Thus, the attempt rate in (3) in the saturated case is given by

$$\theta_{sa} = nq. \quad (7)$$

If a node starts transmission, then it will not stop until a collision occurs. In each slot of the busy period, the probability that at least one node sends request is $1-e^{-\theta_{sa}}$. If other $n-1$ nodes request transmission with probability $1-e^{-\theta_{sa}}$, then a collision occurs in the current time slot, i.e., the busy period ends. Thus the busy period follows a geometric distribution with parameter $1-e^{-\theta_{sa}}$. According to (7), the mean length of busy period in the saturated case can be written as

$$\overline{B}_{sa} = \frac{1}{1-e^{-\theta_{sa}}} = \frac{1}{1-e^{-nq}}. \quad (8)$$

[4]Given the system input parameters such as the traffic input and the transmission probability of HoL packet, how to determine the SAST network is saturated or not remains an unsolved issue and will be one of our future work.

On the other hand, in the unsaturated case, the data queue of the node may be empty. As the number of data packets in the queue is finite, the node might cease the data transmission process when the queue is empty, even if the channel is available. Moreover, the data throughput is equal to the aggregate input rate, i.e., $\lambda_{out}^d = \frac{\overline{B}_{unsa}}{\overline{B}_{unsa}+\overline{V}_c} = n\lambda = \hat{\lambda}$, based on which the mean length of busy period in the unsaturated case can be derived as

$$\overline{B}_{unsa} = \frac{\hat{\lambda}}{1-\hat{\lambda}}\overline{V}_c. \quad (9)$$

Obviously, $\overline{B}_{unsa}$ depends on the mean length of channel vacation period $\overline{V}_c$. According to (6), $\overline{V}_c$ is determined by the probability p_0 and the attempt rate θ, which will be derived in Section V-C.

B. Node Vacation Queuing Analysis

In this subsection, we take a closer look at the behavior of nodes. Intuitively, the queue length in each node is determined by the packet arrival process and packet transmission process.

Let us first discuss the packet arrival process in different periods. Denote A_B as the number of packet arrivals during a busy period. Denote A_V as the number of packet arrivals during a node vacation period. We will give a detailed analysis of the distribution of A_B and A_V, respectively.

1) Distribution of the Number of Arrivals During the Busy Period: When a tagged node enters busy period, the Probability Generating Function (PGF) of A_B can be easily expressed as

$$\begin{aligned}G_{A_B}(z) &= \sum_{j=1}^{\infty}\sum_{i=1}^{j} \Pr\{B=j\}\binom{j}{i}\lambda^i(1-\lambda)^{j-i}z^i \\ &= \sum_{j=1}^{\infty}\Pr\{B=j\}(\lambda z+1-\lambda)^j \\ &= G_B(\lambda z+1-\lambda) \\ &= G_B(1) + \lambda G_B'(1)(z-1) \\ &\quad + \frac{\lambda^2}{2}G_B''(1)(z-1)^2 + \cdots\end{aligned} \quad (10)$$

If the aggregate input rate $\hat{\lambda} = n\lambda$ is fixed, then with a large n, we can deduce that $\lambda^i \approx 0$ where $i \geq 2$. Thus $G_{A_B}(z)$ can be approximated as

$$G_{A_B}(z) \approx 1 - \lambda\overline{B} + \lambda\overline{B}z + o\left(\frac{1}{n}\right). \quad (11)$$

2) Distribution of the Number of Arrivals During the Node Vacation Period: When a busy period B ends, the node shifts to vacation period V. Based on whether the node's buffer is empty or non-empty, we can divide the node vacation into two types:

- **Type-1** vacation V_1: the buffer is non-empty at the beginning of vacation and the tagged node competes for the channel immediately.
- **Type-0** vacation V_0: the buffer is empty at the beginning of vacation and the tagged node will not compete for the channel until a new packet arrives.

Let A_{V1} be the number of packet arrivals in a Type-1 vacation. Lemma 1 presents its PGF.

Lemma 1: The PGF of A_{V1} can be expressed as

$$G_{A_{V1}}(z) = \frac{qe^{-\theta}(1-\lambda+\lambda z)}{1-\left[1-qe^{-\theta}+\left(G_{A_B}(z)-1\right)(1-q)\theta e^{-\theta}\right](1-\lambda+\lambda z)}. \quad (12)$$

Proof: See AppendixA. ∎

Let A_{V0} be the number of packet arrivals in a Type-0 vacation. If the tagged node shifts to a Type-0 period, it stays silent until a new packet arrives and then it competes for the channel. Thus, a Type-0 period consists of two stages: (1) waiting for a new packet; (2) competing for the channel (A_{V1} packets arrive). Thus, we obtain

$$A_{V0} = 1 + A_{V1}. \quad (13)$$

Let $G_{A_V}(z)$ be the PGF of A_V. Combining Lemma 1 and (13), we have Lemma 2.

Lemma 2: When the number of nodes n is large, $G_{A_V}(z)$ can be expressed as

$$G_{A_V}(z) = \begin{cases} \frac{\beta}{1-(1-\beta)z}\left(1-\lambda+\lambda z\right) + o\left(\frac{1}{n}\right), & 1-p_0 \\ \frac{\beta}{1-(1-\beta)z}\left(1-\lambda+\lambda z\right) z + o\left(\frac{1}{n}\right), & p_0 \end{cases} \quad (14)$$

where p_0 is the probability that a transmission node can clear its buffer within one busy period and

$$\beta = \frac{qe^{-\theta}}{qe^{-\theta} + \lambda\left[1 - qe^{-\theta} + \bar{B}\left(1-q\right)\theta e^{-\theta}\right]}. \quad (15)$$

Proof: Substituting (11) into (12), we have $G_{A_{V1}}(z) = \frac{\beta}{1-(1-\beta)z}(1-\lambda+\lambda z) + o\left(\frac{1}{n}\right)$. Moreover, with (13), we have $G_{A_{V0}}(z) = zG_{A_{V1}}(z)$. Note that the probability of one node moving to Type-1 vacation and Type-0 corresponds to p_0 and $1-p_0$, respectively. Then (14) can be obtained by combining $G_{A_{V0}}(z)$ and $G_{A_{V1}}(z)$. ∎

This subsection clearly shows that the attempt rate θ and the probability p_0 are both the key parameters in determining the mean length of busy period $\overline{B}$, node vacation period $\overline{V}$, arrivals during the busy period $\overline{A_B}$ and arrivals during the node vacation period $\overline{A_V}$. In the next subsection, we analyze the attempt rate θ and probability p_0.

C. Attempt Rate

In a node cycle, the buffer of the node is cleared at the end of a busy period with probability p_0. The idle node becomes backlogged when a new packet arrives. As the arrival of data packets follows a Bernoulli process with parameter λ, the average inter-arrival time of the packets is $1/\lambda$ slots. Thus, the average time that the buffer of a node remains empty in a node cycle is p_0/λ slots. Based on that, the probability that the buffer of a node remains empty is given by

$$1 - p_{ne} = \frac{p_0/\lambda}{\overline{C}}. \quad (16)$$

When the network is unsaturated, the mean number of new packets arrived during a node cycle is equal to the mean length of busy period, i.e.,

$$\lambda\overline{C} = \overline{B}. \quad (17)$$

Combining (3), (16) and (17), we have

$$\theta = nq(1 - \frac{p_0}{\overline{B}}). \quad (18)$$

Next, we will discuss the packet transmission process, where p_0 and $\overline{B}$ can be obtained to complete the derivation of the attempt rate θ. Let Q_t and P_t be the queue length at the beginning and the end of the t-th busy period of a node, respectively. When t tends to infinity, the queue length will follow a steady-state distribution, i.e., $Q = \lim_{t\to\infty} Q_t$ and $P = \lim_{t\to\infty} P_t$. Let $G_Q(z)$ and $G_P(z)$ be the PGF of Q and P, respectively. For one hand, given the distribution of Q, we derive the expression of $G_P(z)$ shown in Lemma 3.

Lemma 3: When the number of nodes n is large, given $G_Q(z)$, we have

$$G_P(z) = G_Q\left(e^{-\theta}\right) + \frac{1-e^{-\theta}}{e^{-\theta}-z}\left[G_Q\left(e^{-\theta}\right) - G_Q\left(z\right)\right]. \quad (19)$$

Proof: See Appendix B. ∎

Based on (19), the probability p_0 can be obtained as

$$p_0 = \Pr\left\{P=0\right\} = G_P\left(0\right) = \frac{G_Q\left(e^{-\theta}\right)}{e^{-\theta}}. \quad (20)$$

On the other hand, given the distribution of P, we can derive the expression of $G_Q(z)$. Intuitively, a node experiences a node vacation period before shifting to next busy period. Thus, the queue length at the beginning of the $(t+1)$-th busy period Q_{t+1} is composed of the packets in the buffer at the end of the t-th busy period P_t and the packets newly arrived during the t-th vacation period, denoted by A_{Vt}. Thus, we can obtain $Q_{t+1} = P_t + A_{Vt}$. In the steady state, we have

$$Q = \lim_{t\to\infty} Q_{t+1} = P + A_V. \quad (21)$$

The PGF of Q can be derived as

$$\begin{aligned} G_Q\left(z\right) &= E\left[z^{P+A_V}\right] = E\left[z^{P+A_{V0}}|P=0\right]\Pr\left\{P=0\right\} \\ &+ \sum_{j=1}^{\infty} E\left[z^{P+A_{V1}}|P=j\right]\Pr\left\{P=j\right\} \\ &= p_0 E\left[z^{A_{V0}}\right] + \sum_{j=1}^{\infty} E\left[z^{A_{V1}}\right] E\left[z^{P}|P=j\right]\Pr\left\{P=j\right\} \\ &= p_0 A_{V0}\left(z\right) + \left(G_P\left(z\right) - p_0\right) A_{V1}\left(z\right). \end{aligned} \quad (22)$$

We can see from (19) and (22) that $G_P\left(z\right)$ and $G_Q\left(z\right)$ couple with each other, where each of them is the function of another one. It is necessary to use iterative algorithms to calculate $G_Q(z)$ and $G_P(z)$ and an appropriate initializing point of either Q or P plays a key role in determining the convergence time and output of the iterative algorithm. To reduce the computational complexity, we notice that as the node vacation period V is typically much longer than the busy period B, the number of arrivals during the vacation period A_V constitutes the main body of Q. By ignoring the number of arrivals during the busy period, the explicit expression of $G_Q(z)$ can be approximated such that iterative calculation is no longer needed.

Lemma 4: When the number of nodes n is large, $G_Q(z)$ can be expressed as

$$G_Q(z) \approx \frac{\alpha z}{1-(1-\alpha)z} + o\left(\frac{1}{n}\right), \quad (23)$$

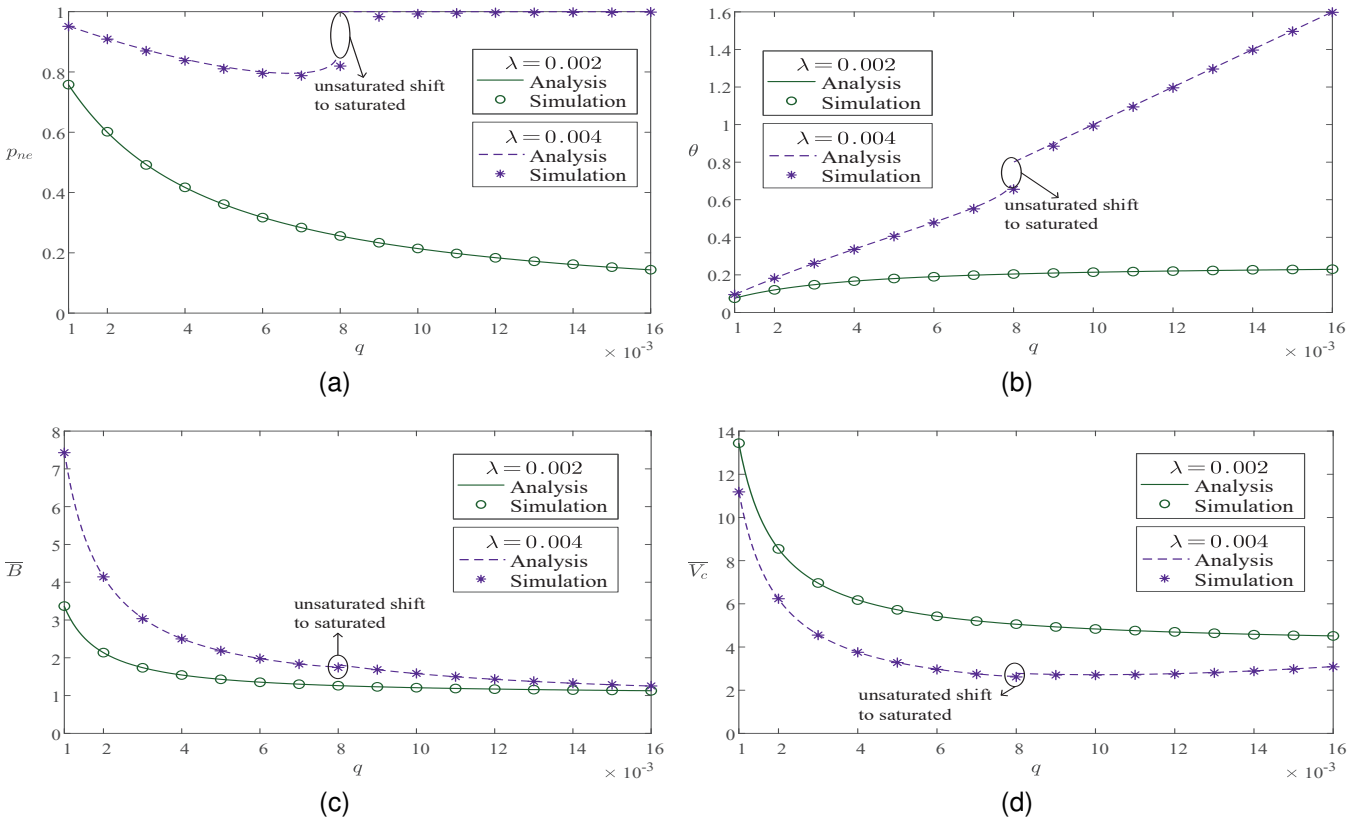

Fig. 3. (a) Non-empty Probability of nodes p_{ne}, (b) Attempt rate θ, (c) Mean length of channel busy period $\overline{B}$, and (d) Mean length of channel vacation period $\overline{V_c}$ versus the transmission probability q. n=100, $\lambda = \hat{\lambda}/n \in \{0.002, 0.004\}$.

where

$$\alpha = 1 - \frac{\theta}{nq}. \tag{24}$$

Proof: See Appendix C. ■

Based on Lemma 4, (20) can be rewritten as

$$p_0 = \frac{1 - \frac{\theta}{nq}}{1 - \frac{\theta e^{-\theta}}{nq}}. \tag{25}$$

Based on (6) and (25), (9) can be rewritten as

$$\overline{B} = \frac{\hat{\lambda}}{1-\hat{\lambda}}\overline{V_c} = \frac{\hat{\lambda}}{1-\hat{\lambda}}\left(\frac{e^{\theta}}{\theta} - \frac{1 - \frac{\theta}{nq}}{1 - \frac{\theta e^{-\theta}}{nq}}\right). \tag{26}$$

With a large n, we can get the fixed-point equation of the attempt rate θ by substituting (25) and (26) into (18) as follows:

$$\frac{e^{\theta}}{\theta} + \frac{\theta}{nq} - \frac{1}{nq} = \frac{1}{\hat{\lambda}}. \tag{27}$$

Fig. 3 demonstrates how the attempt rate θ, the probability of non-empty buffer of nodes p_{ne}, the mean length of busy period $\overline{B}$, and the mean length of channel vacation period $\overline{V_c}$ vary with the transmission probability q for $n = 100$ with the node input rate $\lambda = 0.002$ or 0.004 (i.e., the aggregate input rate $\hat{\lambda} = 0.2$ or 0.4). Simulations follow Section III and each case runs for 10^7 time slots via a MATLAB-based simulator.

In a unsaturated case(with a light node input rate $\lambda = 0.002$), as the transmission probability q increases, the attempt rate θ increases, but p_{ne}, $\overline{B}$ and $\overline{V_c}$ decreases. On the other hand, with a heavy node input rate of $\lambda = 0.004$, there exists a stage exhibiting a jump change, implying that the network transitions from an unsaturated to a saturated state. As illustrated in Fig. 3a, due to severe network congestion resulting from the increasing transmission probability q, the number of successfully transmitted packets gradually becomes less than the newly arriving packets over time. Consequently, the non-empty probability of nodes p_{ne} rapidly increases to 1, and the network enters a saturated state. In this scenario, the attempt rate is determined by (7), rather than (27), as depicted in Fig.3b. The shift in the mean busy period $\overline{B}$ and channel vacation period $\overline{V_c}$ can be observed in Figs. 3c and 3d, respectively. However, this change is not significant.

VI. Throughput Analysis

This section focuses on the optimal throughput performance of SAST. To push the throughput performance to the limit, we consider the saturated case, where each node always has packets to send[5], i.e., $p_0 = 0$ and $p_{ne} = 1$. By combining (6) and (7), the mean length of channel vacation period in the saturated case becomes

$$\overline{V_{c,sa}} = \frac{e^{nq}}{nq}. \tag{28}$$

According to (8) and (28), the access throughput and the data throughput in the saturated case can be written as

$$\lambda_{out}^{a} = \frac{nq\left(1 - e^{-nq}\right)}{e^{nq} + nq - 1} \tag{29}$$

[5]By regarding an n-node SAST network as an n-queue-single-server system, we can see that the network throughput $\hat{\lambda}_{out}$ is indeed the system output rate, which is equal to the aggregate input rate $\hat{\lambda}$ if $p_{ne} < 1$ [48]. As we are interested in the throughput performance limit of SAST and how to achieve it, only the saturated case is considered in this section.

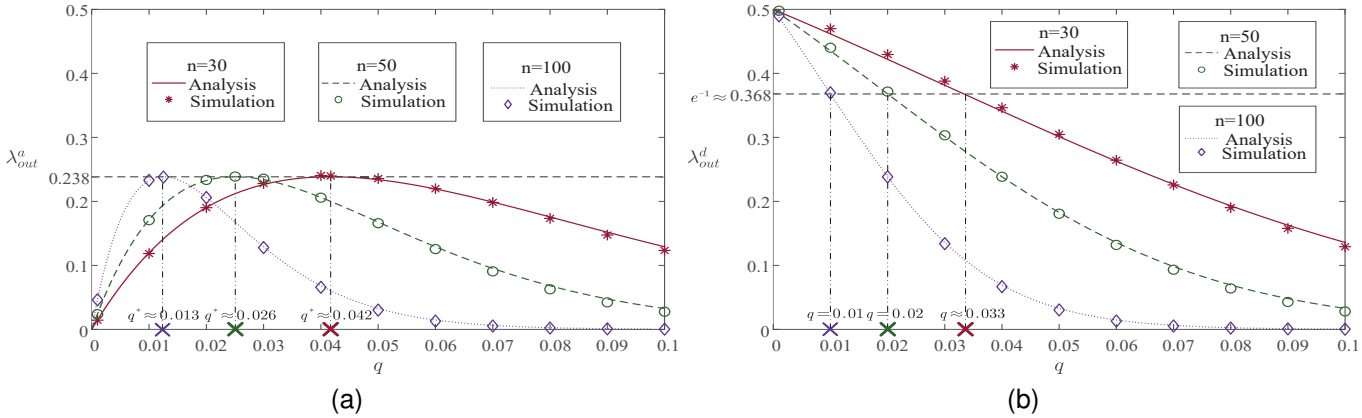

Fig. 4. (a) Access throughput λ^a_{out} versus the transmission probability q. (b) Data throughput λ^d_{out} versus q. n=100, $\lambda = \hat{\lambda}/n \in \{0.002, 0.004\}$.

and

$$\lambda^d_{out} = \frac{nq}{e^{nq} + nq - 1}. \tag{30}$$

Theorem 1 shows the maximum access throughput λ^a_{max} and corresponding optimal transmission probability q^* in the saturated case. Theorem 2 shows the maximum data throughput λ^d_{max} is 0.5.

Theorem 1: When the network is saturated, the maximum access throughput is given by

$$\lambda^a_{max} \approx 0.2384, \tag{31}$$

which is achieved if and only if

$$q^* \approx 1.2515/n. \tag{32}$$

Theorem 2: When the network is saturated, the data throughput is a monotonic decreasing function of nq, and the maximum data throughput is given by

$$\lambda^d_{max} = \lim_{nq \to 0} \lambda^d_{out} = \lim_{nq \to 0} \frac{nq}{e^{nq} + nq - 1} = \frac{1}{2}. \tag{33}$$

The proofs of Theorem 1 - 2 are presented in Appendix D.

Fig. 4 depicts how the access throughput λ^a_{out} and data throughput λ^d_{out} vary with the transmission probability q in saturated case with the number of node $n \in \{30, 50, 100\}$. It can be seen in Fig. 4a that when the transmission probability q is small, the access throughput λ^a_{out} increases as q increases because more and more nodes can access to the network and the contention is not serious as well. But if q is large, the access throughput λ^a_{max} decreases because of the mounting channel contention. The maximum access throughput λ^a_{max} can be achieved when the transmission probability q is tuned properly, i.e., $q = q^* \approx 1.2515/n$. In addition, the maximum access throughput $\lambda^a_{max} \approx 0.2384$ is not affected by the number of nodes. As for the data throughput λ^d_{out} in saturated case in Fig. 4b, the analysis and simulation both show that λ^d_{out} decreases as nq increases. Taking the maximum throughput in classic slotted Aloha $1/e$ as the threshold, (30) tells us that with $nq < 1$, the data throughput λ^d_{out} in the saturated case will always be larger than $1/e$ and smaller than 0.5. Considering the access throughput λ^a_{out} and data throughput λ^d_{out} jointly, it can be found that the SAST scheme achieves high data throughput with a low access throughput. Intuitively, when the transmission probability q is small, it is difficult for the backlogged nodes to access the channel and transmit packets. Once an access request is successful, a large number of packets will be transmitted with a high successful probability, indicating that although the system access throughput is small, a large data throughput can be achieved with a small q.

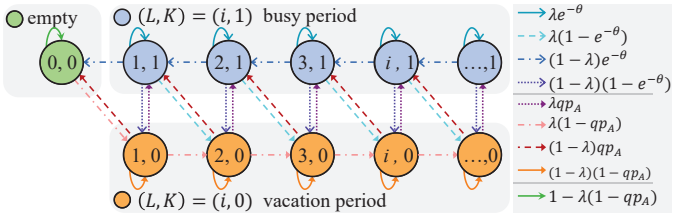

Fig. 5. Markov chain (L_t, K_t) of the state transition of node queue length at slot t.

VII. Delay Analysis

This section derives the mean access delay D_A, the mean packet delay D_P and demonstrates how the delay performance varies with the transmission probability q.

A. Access Delay D_A

Recall that the access delay is defined as the long-term mean time length of access requests (i.e., the HoL packet) from generation to acceptance, which is essentially equivalent to the definition of Type-1 vacation V_1 in Section V-B, i.e., $D_A = \overline{V_1}$. Accordingly, by combining $\overline{A_{V_1}} = \lambda \overline{V_1}$, (11) and (12), the mean access delay is given by

$$D_A = \overline{V_1} = \frac{\overline{A_{V_1}}}{\lambda} = \frac{\frac{\theta}{nq}}{\lambda\left(1 - \frac{\theta}{nq}e^{-\theta}\right)}. \tag{34}$$

When the network is saturated, the access throughput of single node is given by $\frac{\lambda^a_{out}}{n} = \frac{1}{\overline{B}_{sa} + \overline{V}_{1,sa}} = \frac{q\left(1-e^{-nq}\right)}{nq+e^{nq}-1}$, based on which we can further obtain the explicit expression of the mean access delay $D_{A,sa}$, i.e.,

$$D_{A,sa} = \overline{V}_{1,sa} = \frac{e^{nq} + nq - 1 - q}{q\left(1 - e^{-nq}\right)} \tag{35}$$

by combining (8).

B. Packet Delay D_P

We introduce a two-dimensional Markov chain (L_t, K_t) to characterize the behavior of the node's queue, where $L_t \in \{0, 1, \cdots\}$ denotes the queue length of a tagged node at slot t, and $K_t = 1$ or 0 indicates that the tagged node is in busy period or vacation period, respectively. In each slot, the new packet (at most one packet) arrives in the queue with probability λ, while at most one packet may leave due to successful transmission. Fig. 5 shows the state transition process of each individual node, where p_A denotes the probability of successful transmission of access requests. The steady-state probability distribution of the Markov chain in Fig. 5 can be obtained as

$$\begin{cases} \pi_{10} = \left(\frac{1}{1-\frac{\lambda}{1-\lambda}\gamma} + \frac{\lambda}{1-\lambda} - 1 + \frac{1}{1-\gamma} + \frac{(1-\lambda)p_A + \lambda e^{-\theta}}{\lambda(1-p_A)} \right)^{-1} \\ \pi_{00} = \frac{(1-\lambda)p_A + \lambda e^{-\theta}}{\lambda(1-p_A)} \pi_{10} \\ \pi_{11} = \frac{\lambda}{1-\lambda} \pi_{10} \\ \pi_{i0} = \gamma^{i-1} \pi_{10}, i \geq 2 \\ \pi_{i1} = \frac{\lambda}{1-\lambda} \pi_{i0} = \left(\frac{\lambda}{1-\lambda} \gamma \right)^{i-1} \pi_{10}, i \geq 2 \end{cases}, \tag{36}$$

where $\gamma = \frac{\lambda\left(\lambda\left(1-e^{\theta}\right)+(1-\lambda)(1-p_A)\right)}{(1-\lambda)(\lambda e^{-\theta}+(1-\lambda)p_A)}$. Note that $E[\pi] = \sum_{i=1}^{\infty} (\pi_{i0} + \pi_{i1})\, i$ is the mean length of node buffer queue $\overline{L}$.

Now we discuss the probability of successful transmission of access requests p_A. The channel is available for transmission if and only if the channel is in vacation period with probability $1 - \lambda_{out}^d$ or in the last time slot of busy period in which the node finished the transmission of the last packet. For the latter case, it only occurs when the transmitting node clears its buffer in this busy period with probability p_0, and each channel cycle has at most one such slot, i.e., the probability of the latter case is given by $\frac{p_0}{C_c} = \frac{p_0}{B+V_c}$. In addition to the channel, the access probability of other nodes $e^{-\theta}$ is also taken into consideration. According to (6), (25) and (26), the probability of successful transmission of access requests p_A is given by

$$p_A = \left(1 - \frac{\lambda\theta}{q}\right) e^{-\theta}. \tag{37}$$

By combining (36) and (37), the mean length of node buffer queue $\overline{L}$ is given by

$$\overline{L} = \frac{\frac{1}{1-\lambda} + \frac{\gamma(2-\gamma)}{(1-\gamma)^2} + \frac{\frac{\lambda\gamma}{1-\lambda}\left(2-\frac{\lambda\gamma}{1-\lambda}\right)}{\left(1-\frac{\lambda\gamma}{1-\lambda}\right)^2}}{\frac{1}{1-\frac{\lambda\gamma}{1-\lambda}} + \frac{\lambda}{1-\lambda} - 1 + \frac{1}{1-\lambda} + \frac{(1-\lambda)p_A + \lambda e^{-\theta}}{\lambda(1-p_A)}}. \tag{38}$$

According to the Little's law, the mean packet delay can be derived as

$$D_P = \frac{\frac{1}{1-\lambda} + \frac{\gamma(2-\gamma)}{(1-\gamma)^2} + \frac{\frac{\lambda\gamma}{1-\lambda}\left(2-\frac{\lambda\gamma}{1-\lambda}\right)}{\left(1-\frac{\lambda\gamma}{1-\lambda}\right)^2}}{\lambda\left(\frac{1}{1-\frac{\lambda\gamma}{1-\lambda}} + \frac{\lambda}{1-\lambda} - 1 + \frac{1}{1-\lambda} + \frac{(1-\lambda)p_A + \lambda e^{-\theta}}{\lambda(1-p_A)}\right)}. \tag{39}$$

Fig. 6 depicts how the mean access delay D_A and the mean packet delay D_P vary with the transmission probability q. When the network is unsaturated, it can be seen from Figs. 6a and 6b that as the transmission probability q increases, the mean access delay D_A and the mean packet delay D_P both decrease. This is because when the channel contention is light, increasing the transmission probability provides nodes with more opportunities to secure channel availability, thereby achieving timely packet delivery. However, when the network is saturated, the mean access delay $D_{A,sa}$ first decreases and then increases, which is minimized when q is set according to (32). Here, increasing q will intensify the congestion and bring high access delay. It can be observed from Fig. 6c and Fig. 4a that trends of the mean access delay and access throughput in the saturated case are exactly opposite.

VIII. Case Study: 5G Cellular Network

In this section, we explain how the proposed scheme can be used in current 5G cellular network and compare its performance with 2-step SDT scheme by leveraging the signaling-to-throughput Ratio (STR), i.e., the signaling overhead per successful data packet per slot.

A. 2-step SDT Scheme

According to 3GPP specifications [5], with 2-step SDT scheme, each node transmits its small packets in the random access procedure, such that the connection with base station is no longer needed. As shown in Fig. 7a, each backlogged node transmits one data packet along with a preamble in MsgA. If the receiver replies with the Random Access Response (RAR) and the Contention Resolution Response in MsgB, then the node knows whether its MsgA transmission is successful or not. Although the signaling overhead for connection establishment is avoided, the signaling overhead due to failed requests still exists and may increase when the number of nodes is large.

To characterize the signaling overhead in details, denote s (in a unit of bits) as the average size of signaling message exchanged between node and receiver. To be specific, one MsgA or MsgB consists of an average of s bits of signaling. Denote S as the time-average amount of signaling overhead per time slot. Denote F as the time-average amount of failed access per times slot. Denote STR as the ratio of average signaling overhead per time slot to the data throughput, i.e., signaling overhead per successful data packet per slot.

By observing Fig. 7a, we can see that no matter the access request is successful or not, two signaling messages are required. Thus, we have the signaling overhead per time slot in 2-step SDT scheme as $S_{SDT} = 2sF + 2s\lambda_{out}^a = 2s\frac{\lambda_{out}^a}{p_A}$. On the other hand, the data throughput is equal to the access throughput, i.e., $\lambda_{out}^d = \lambda_{out}^a$. The STR in 2-step SDT scheme can be obtained as

$$\text{STR}_{SDT} = \frac{S_{SDT}}{\lambda_{out}^d} = \frac{2s(F + \lambda_{out}^a)}{\lambda_{out}^a} = \frac{2s\lambda_{out}^a}{\lambda_{out}^a p_A} = \frac{2s}{p_A}. \tag{40}$$

B. Slotted Aloha with Successive Transmission Scheme

As shown in Fig. 7b, similar to 2-step SDT scheme, each node with SAST will experience the same step for the transmission of the HoL packet. Upon the HoL packet is transmitted successfully, the node only needs to transmit

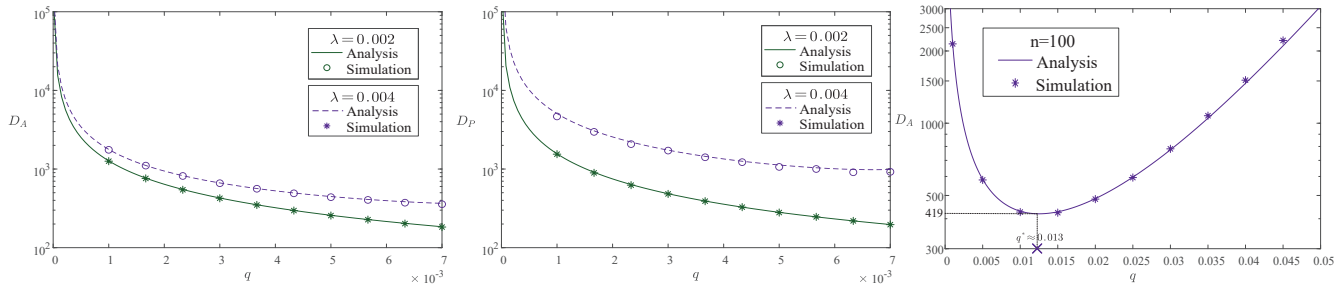

Fig. 6. (a) Mean access delay $D_{A,sa}$ versus the transmission probability q in unsaturated case. (b) Mean packet delay D_P versus q in unsaturated case. (c) Mean access delay D_A versus q in saturated case. n=100. $\lambda = \hat{\lambda}/n \in \{0.002, 0.004\}$.

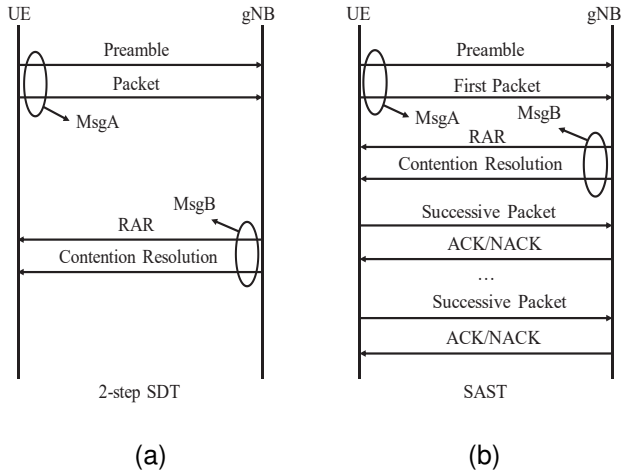

Fig. 7. Interaction between node and base station of (a) 2-step SDT random access, (b)Slotted Aloha with successive transmission random access.

the next packet without the preamble. If the receiver replies with an ACK message, then the node can transmit another packet successively. On the contrary, if a NACK message is replied, then the node has to repeat the transmission process. Here, ACK/NACK message or the subsequent failed packet are regarded as signaling overhead.

As one successful access in SAST scheme can transmit $\overline{B}$ packets (including the HoL packet), two cases may occur: (1) the transmission is not interrupted with probability p_0, which requires $\overline{B}-1$ signaling overheads; (2) the transmission is interrupted with probability $1-p_0$, which requires $\overline{B}+1$ signaling overheads (extra 2 signaling overheads are from the last failed packet). Thus, we have the signaling overhead per time slot in SAST scheme as $S_{SAST} = 2sF + \lambda^a_{out}\left[2s + p_0\left(\overline{B}-1\right)s + \left(1-p_0\right)\left(\overline{B}+1\right)s\right] = \lambda^a_{out}s\left(\frac{2}{p_A} + \overline{B} + 1 - 2p_0\right)$. The STR in SAST scheme can be obtained as

$$\mathrm{STR}_{SAST} = \frac{S_{SAST}}{\lambda^d_{out}} = s + s\frac{1}{\overline{B}}\left(\frac{2}{p_A} + 1 - 2p_0\right). \quad (41)$$

When the network is saturated, (41) can be further derived as

$$STR_{SAST,sa} = s\left(2e^{nq} + 2nq - e^{-nq}\right), \quad (42)$$

in which the STR is a monotonically increasing function of q.

C. Performance Comparison

Fig. 8a depicts how the STR varies with the transmission probability q in saturated case. It can be seen that when the transmission probability[6] $q < \frac{\mathbb{W}_0(0.5)}{n} = 0.0035$, the SAST scheme outperforms 2-step SDT scheme from the perspective of STR (nearly half at most better in the case of $q \to 0$). In particular, when $q = 0.0035$, 2-step SDT and SAST have the same STR performance while according to (30), the data throughput of SAST is $\lambda^d_{out} = 0.4549$, still higher than e^{-1}, i.e., the maximum data throughput of 2-step SDT scheme. Fig. 8b demonstrates how the STR varies with the aggregate input rate $\hat{\lambda}$ in unsaturated case. It can be seen that the SAST scheme always outperforms 2-step SDT scheme. To be specific, given the aggregated input rate $\hat{\lambda} = e^{-1}$, the STR of the SAST scheme is only 30% of that of 2-step SDT scheme.

IX. Discussion

There are also a few key assumptions that may be relaxed when extending the analysis to a variety of wireless communication systems for SAST:

a) *Multi-packet reception (MPR)*: In this paper, we adopt the classic collision model assumption, where simultaneous transmissions of two or more packets leads to decoding failure. However, in practical network, to enhance the network performance, the access point could adopt advanced receiver structure, e.g., capture mode, such that multiple packets could be successfully decoded within a single time slot. With MRP, it has been observed that the successful decoding probability of data packets is influenced not only by the number of concurrent transmissions but also by the encoding rate and transmission power of each packet [50]. Given MPR capabilities at the receiver, optimizing the performance of SAST necessitates the joint selection of transmission probabilities alongside other system parameters, including encoding rates and transmission powers. This optimization framework requires further investigation.

b) *Channel fading*: As collision model is assumed, a packet can be successfully transmitted as long as no other concurrent transmissions. In practical wireless communication systems, the effect of channel fading on data packet/feedback (ACK/NACK) decoding always exists. It can be expected that

[6]The root of the fixed-point equation $STR_{SAST,sa} = \mathrm{STR}_{SDT}$ is 0.0035, which is the intersection point in Fig. 8a. $\mathbb{W}_0$ represents one of the two branches of Lambert W function, i.e., $z = \mathbb{W}_0(z)e^{\mathbb{W}_0(z)}$ for any complex number z [49].

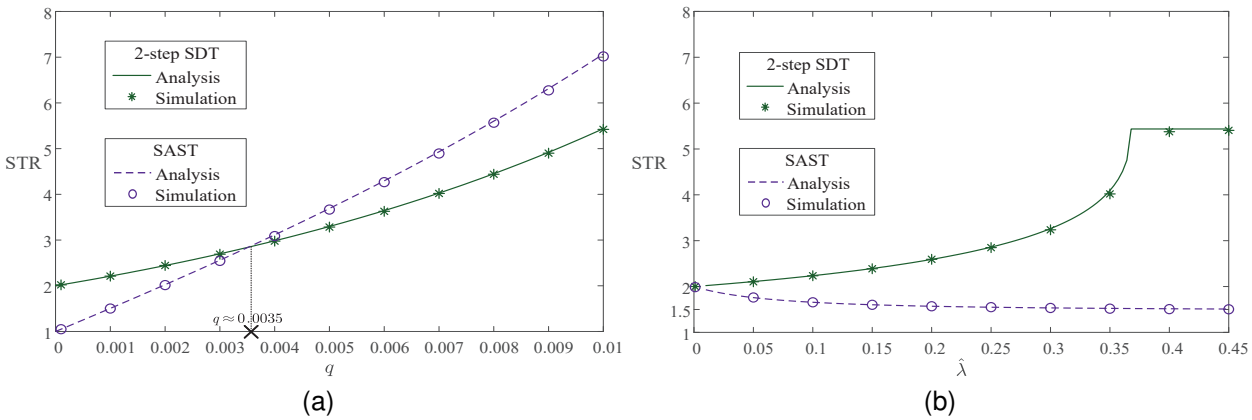

Fig. 8. (a) STR versus the transmission probability q in saturated case. (b) STR versus the aggregate input rate $\hat{\lambda}$ in unsaturated case. q in 2-step SDT and SAST scheme is set to 0.01 and 0.001, respectively. n=100. s=1.

the mean length of channel vacation period will be stretched while that of channel busy period will be shortened. By following the same methodology in this paper, the analysis can be extended to incorporate the effect of channel fading.

c) *Energy efficiency*: Energy efficiency is a commonly used metric in existing literature, which is usually defined as the ratio of the number of successfully transmitted information bits to the total energy consumption [51]. By comparing the definition of energy efficiency and that of STR, it is clear that a larger STR indicates that more energy are used to transmit signaling bits rather than information bits, resulting in a lower energy efficiency. Therefore, STR is a bellwether of energy efficiency. The analysis in this paper can be extended to further consider the energy efficiency. The key to this extension lies in using the two-dimensional Markov chain (L_t, K_t) to characterize the behavior of the node's queue and then derive the average energy consumption of each node in each time slot.

d) *Heterogeneous scenario*: This paper considers the homogeneous scenario, where all nodes have the same data input rate and transmission probability. The analysis can be extended to the heterogeneous scenario, where nodes have different input rates and transmission probabilities. Specifically, nodes can be divided into groups depending on the system input parameters, where nodes in the same group have the same configuration while the configurations of nodes differ from group to group. By characterizing the state transition of queue length of each node in different groups and the aggregate activities on channel contention, vacation queue analysis could be extended to heterogeneous scenarios to exploit the performance limit of SAST.

e) *Multi-channel scenario*: This paper focuses on the single-channel scenario, where all nodes share a single channel. To extend the proposed scheme to the multi-channel scenario, where backlogged nodes transmit their packets over a randomly selected channel, the number of nodes contending in each channel can be approximated by the ratio of the total number of nodes to the number of channels. Based on this assumption, the analysis presented in this paper can be readily extended to multi-channel scenarios.

X. Conclusion

This paper proposes the SAST scheme for boosting the throughput and the signaling-to-Throughput Ratio (STR) performance of slotted Aloha. By establishing vacation queueing models for characterizing the behavior of both node and channel, the access throughput and data throughput are derived as functions of system parameters and are further optimized by properly tuning the transmission probability. To evaluate the delay performance of SAST scheme, a two-dimensional Markov chain is formulated for each node, based on which the access delay and packet delay are analyzed. To illustrate the practical insights of the SAST scheme, the 2-step SDT scheme in 5G cellular network is further considered as benchmark for comparison, where STR of both schemes are derived.

The analysis demonstrates that the maximum data throughput of the SAST scheme can reach up to 0.5, surpassing the performance bottleneck of classic slotted Aloha, i.e., $1/e$. In addition, achieving such optimum throughput performance of SAST is simple: Reducing the transmission probability of each node if the input rate is large enough. In other words, the optimal setting of the transmission probability is independent of the system input parameters, which facilitates its implementation in practical system. Moreover, the 5G case study reveals that compared with 2-step SDT scheme, the proposed SAST scheme can achieve a better throughput performance with much lower signaling overhead, indicating the SAST scheme is promising to be used in practical 5G for supporting a broad spectrum of IoT applications with stringent requirement on throughput and energy efficiency.

Appendix A
Proof of Lemma2

Three cases may occur at the first slot of the type-1 vacation:

C1: With probability $qe^{-\theta}$, the tagged node transmits packets to the receiver and other $n-1$ nodes do not request transmission. In this case, the node vacation period only holds one slot. As the number of packet arrivals in one slot follows Bernoulli distribution with parameter λ whose PGF is $I(z) = 1 - \lambda + \lambda z$, the PGF of A_{V1} in Case 1 is

$$G_{A_{V1}}(z)|\mathrm{C1} = 1 - \lambda + \lambda z. \tag{43}$$

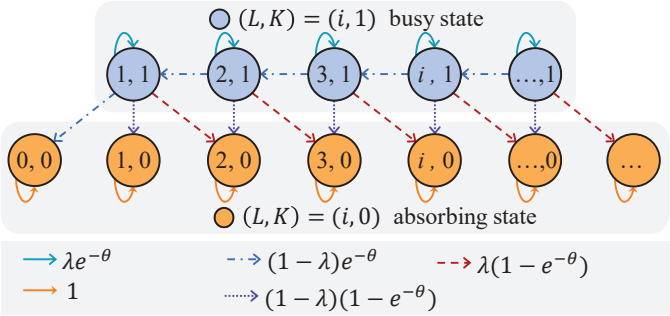

Fig. 9. Absorbing Markov chain (L_t,K_t) of the state transition of queue length when entering busy period.

C2: With probability $(1-q)\theta e^{-\theta}$, the tagged node does not transmit with probability $1-q$, and only one of the other $n-1$ nodes transmits successfully. The tagged node remains in vacation period. In this case, the tagged node will compete for the channel until it succeeds entering its next busy period. Before that, it will experience three stages in sequence: (1) one slot that it does not compete; (2) a busy period of another node; (3) a new type-1 vacation. Accordingly, the PGF of A_{V1} in Case 2 is

$$G_{A_{V1}}(z)|\mathrm{C2}=(1-\lambda+\lambda z)\times G_{A_B}(z)\times G_{A_{V1}}(z)\,. \quad (44)$$

C3: With probability $1-(1-q)\theta e^{-\theta}-qe^{-\theta}$, no one succeeds in the first slot and the tagged node has to compete for the channel at the next time slot. Two stages will be experienced: (1) one free slot; (2) a type-1 vacation period. Thus, the PGF of A_{V1} in Case 3 is

$$G_{A_{V1}}(z)|\mathrm{C3}=(1-\lambda+\lambda z)\times G_{A_{V1}}(z)\,. \quad (45)$$

According to the law of total probability, (12) can be obtained by combining the above three cases.

Appendix B
Proof of Lemma 3

In this appendix, we first derive the distribution and PGF of P given Q, and then derive the PGF $G_P(z)$ given $G_Q(z)$. To derive the distribution of P, let us define an absorbing Markov Chain (L_t,K_t), where $L_t\in[0,1,\cdots]$ denotes the queue length of nodes when entering busy period at the t-th transmission, and $K_t\in\{0,1\}$ denotes the busy period available at the t-th transmission. Specifically, when $K_t=1$ or 0, the tagged node can be said in busy period or vacation period, respectively. As shown in Fig. 9, the sates $(L_t,0)$ are absorbing states as they are impossible to leave (i.e., $p_{ii}=1$). And the states $(L_t,1)$ are transient states as from any of these it is possible to reach the absorbing state. When the chain shifts to the state $(L_t,0)$, it is said the busy period has ended, i,e. the limiting probability of absorbing state is the probability of the queue length at the end of busy period P.

Note that the first time slot in a busy period is guaranteed to succeed. The queue length of nodes that enter the absorbing Markov chain are smaller than that at the beginning of their busy period by 1, i.e., $L_t=Q_t-1$. Thus, we can easily obtain that when $Q_t=1$, it goes directly to state $(0,0)$ and $P_t=0$ with probability 1. Except for this situation, those nodes that enter the busy period with $Q_t\geq 2$ can be considered as initialing to enter the absorbing Markov chain with state $(L_t,1)$. By jointly analyzing the two processes of new packet arrival and packet transmission in the busy period, for the current state $(L,K)=(j,1)$, it can be known that:

(1) with probability $\lambda e^{-\theta}$, it has a arrival and transmits a packet successfully, i.e., $(j,1)$ remains in $(j,1)$;
(2) with probability $(1-\lambda)e^{-\theta}$, it does not have a arrival and transmits a packet successfully, i.e., $(j,1)$ transfers to $(j-1,1)$;
(3) with probability $\lambda(1-e^{-\theta})$, it has a arrival and fails to transmit a packet, i.e., the busy period ends and $(j,1)$ transfers to $(j+1,0)$;
(4) with probability $(1-\lambda)(1-e^{-\theta})$, it does not have a arrival and fails to transmit a packet, i.e., the busy period ends and $(j,1)$ transfers to $(j,0)$.

For convenience, we replace λ, $e^{-\theta}$, $1-\lambda$, $1-e^{-\theta}$ with a,b,c,d, respectively. The state transition probability matrix can be written as

$$\begin{pmatrix} \mathbf{T}_{j\times j} & \mathbf{R}_{j\times(j+1)} \\ \mathbf{0}_{(j+1)\times j} & \mathbf{I}_{(j+1)\times(j+1)} \end{pmatrix}_{(2j+1)\times(2j+1)} \quad (46)$$

where the submatrix in the upper left corner, denoted as $\mathbf{T}$, is a j-by-j matrix that represents the transition from the transient states to the transient states, the submatrix in the upper right corner, denoted as $\mathbf{R}$, is a j-by-$(j+1)$ matrix that represents the transition from the transient states to the absorbing states, the submatrix in the lower left corner is a $(j+1)$-by-j zero matrix $\mathbf{0}$ and the submatrix in the lower right corner is a $(j+1)$-by-$(j+1)$ identity matrix $\mathbf{I}$. The first j states are transient and the second $j+1$ states are absorbing (Here j goes into infinity).

Let b_{ij} be the probability that an absorbing chain will be absorbed in the absorbing state s_j if it starts in the transient state s_i. Let $\mathbf{B}$ be the matrix with entries b_{ij}. Then $\mathbf{B}$ is an j-by-$(j+1)$ matrix, and [52]

$$\mathbf{B}=(\mathbf{I}_{j\times j}-\mathbf{T}_{j\times j})^{-1}\mathbf{R}_{j\times(j+1)}. \quad (47)$$

By solving Matrix B, b_{ij} can be obtained as

- $i=1$. $b_{ij}=\begin{cases}\frac{(1-\lambda)e^{-\theta}}{1-\lambda e^{-\theta}}, j=0\\ \frac{(1-\lambda)\left(1-e^{-\theta}\right)}{1-\lambda e^{-\theta}}, j=1\\ \frac{\lambda\left(1-e^{-\theta}\right)}{1-\lambda e^{-\theta}}, j=2\end{cases}$.
- $i\geq 2$. $b_{ij}=$
$$\begin{cases}\left[\frac{(1-\lambda)e^{-\theta}}{1-\lambda e^{-\theta}}\right]^i, j=0\\ \frac{(1-\lambda)\left(1-e^{-\theta}\right)}{1-\lambda e^{-\theta}}\left[\frac{(1-\lambda)e^{-\theta}}{1-\lambda e^{-\theta}}\right]^{i-1}, j=1\\ \frac{(1-\lambda)\left(1-e^{-\theta}\right)}{[1-\lambda e^{-\theta}]^2}\left[\frac{(1-\lambda)e^{-\theta}}{1-\lambda e^{-\theta}}\right]^{i-j}, 2\leq j\leq i\\ \frac{\lambda\left(1-e^{-\theta}\right)}{1-\lambda e^{-\theta}},\ j=i+1\end{cases}.$$

b_{ij} is also the probability of the queue length at the end of busy period except the situation when $Q_t=1$. As $L_t=Q_t-1$, combining the expression of b_{ij}, the distribution of P is then given by

- $Q=1$. $\Pr\{P=0\}=1$.

- $Q=2$. $\Pr\{P=i|Q=2\}=\begin{cases}\frac{(1-\lambda)e^{-\theta}}{1-\lambda e^{-\theta}}, i=0\\ \frac{(1-\lambda)\left(1-e^{-\theta}\right)}{1-\lambda e^{-\theta}}, i=1\\ \frac{\lambda\left(1-e^{-\theta}\right)}{1-\lambda e^{-\theta}}, i=2\end{cases}$.
- $Q\geq 3$. $\Pr\{P=i|Q=j\geq 3\}=$
$$\begin{cases}\left[\frac{(1-\lambda)e^{-\theta}}{1-\lambda e^{-\theta}}\right]^{j-1}, i=0\\ \frac{(1-\lambda)\left(1-e^{-\theta}\right)}{1-\lambda e^{-\theta}}\left[\frac{(1-\lambda)e^{-\theta}}{1-\lambda e^{-\theta}}\right]^{j-2}, i=1\\ \frac{(1-\lambda)\left(1-e^{-\theta}\right)}{\left[1-\lambda e^{-\theta}\right]^2}\left[\frac{(1-\lambda)e^{-\theta}}{1-\lambda e^{-\theta}}\right]^{j-i-1}, 2\leq i\leq j-1\\ \frac{\lambda\left(1-e^{-\theta}\right)}{1-\lambda e^{-\theta}},\ i=j\end{cases}.$$

With a large n , the input rate of a node $\lambda=\hat{\lambda}/n\to 0$. the distribution of P can be further approximated as

$$\Pr\{P=i\}=\begin{cases}(1-e^{-G})e^{-G(Q-i-1)}, i=1,...,Q-1\\ e^{-G(Q-1)}, i=0\end{cases}. \quad (48)$$

Based on (48), (19) in Lemma 3 can be also derived.

APPENDIX C
PROOF OF LEMMA 4

In this appendix, we prove that Q is geometrically distributed when the number of nodes n is large. Assume that Q follows a geometric distribution with parameter α, that is $q_j=\alpha(1-\alpha)^{j-1}, j=1,2,\cdots,$, where $\alpha\in(0,1)$. Thus, the PGF of Q is

$$G_Q(z)=\frac{\alpha z}{1-(1-\alpha)z}+o\left(\frac{1}{n}\right). \quad (49)$$

Substituting (49) into (19) and (20), we have

$$G_P(z)=\frac{1-(1-\alpha)e^{-\theta}z}{[1-(1-\alpha)z]\left[1-(1-\alpha)e^{-\theta}\right]}, \quad (50)$$

and

$$p_0=G_P(0)=\frac{\alpha}{1-(1-\alpha)e^{-\theta}}. \quad (51)$$

By substituting (50), (51) and (14) into the right-hand side of (22), we obtain

$$\begin{aligned}&p_0A_{V0}(z)+[G_P(z)-p_0]A_{V1}(z)\\ &=[G_P(z)-p_0+p_0z]A_{V1}(z)+o\left(\frac{1}{n}\right)\\ &=\left\{\frac{\alpha\left(1-(1-\alpha)e^{-\theta}z\right)}{[1-(1-\alpha)z]\left[1-(1-\alpha)e^{-\theta}\right]}\right.\\ &\quad\left.+\frac{\alpha}{\left[1-(1-\alpha)e^{-\theta}\right]}(z-1)\right\}\\ &\quad\times\frac{\beta}{1-(1-\beta)z}(1-\lambda+\lambda z)+o\left(\frac{1}{n}\right).\end{aligned} \quad (52)$$

In the following, we show that β can be expressed as a function of α. Recall that $Q=P+A_V$, as (21) shows. Thus, $\overline{Q}$ satisfies

$$\overline{Q}=\overline{P}+\overline{A_V}=\overline{P}+p_0\overline{A_{V0}}+(1-p_0)\overline{A_{V1}}, \quad (53)$$

where $\overline{P}=G'_P(1)=G'_Q(1)-\frac{G_Q\left(e^{-\theta}\right)-G_Q(1)}{e^{-\theta}-1}$, $\overline{A_{V1}}=G'_{AV1}(1)=\frac{1}{\beta}+\lambda-1$, and $\overline{A_{V0}}=1+\overline{A_{V1}}=\frac{1}{\beta}+\lambda$. According to (49), the mean length at the beginning of busy period $\overline{Q}$ can be also obtained as $\overline{Q}=\frac{1}{\alpha}$.

Combining with the above derivation, the parameter β can be expressed as a function of α as follows:

$$\beta=\frac{1}{1-\lambda+\frac{1-\alpha}{1-(1-\alpha)e^{-\theta}}} \quad (54)$$

So we can simplified (52) again in terms of the input rate of a node $\lambda=\hat{\lambda}/n\to 0$ as

$$\begin{aligned}&p_0A_{V0}(z)+[G_P(z)-p_0]A_{V1}(z)\\ &=\left\{\frac{\alpha z\left[1-(1-\alpha)\left(e^{-\theta}+1-1\right)\right]}{[1-(1-\alpha)z]\left[1-(1-\alpha)e^{-\theta}\right]}\right\}\\ &\quad\times\frac{\frac{1}{1+\frac{1-\alpha}{1-(1-\alpha)e^{-\theta}}}}{1-\left(1-\frac{1}{1+\frac{1-\alpha}{1-(1-\alpha)e^{-\theta}}}\right)z}\\ &=\frac{\alpha z}{1-(1-\alpha)z}\\ &=Q(z).\end{aligned} \quad (55)$$

This result indicates that Lemma 4 is established.

In the following, we derive the value of parameter α. Recall that the the number of packets transmitted in a busy period is equal to the length of busy period. From an expectation point of view, within a busy period length, the packets in the node queue should satisfy a quantity relationship: $\overline{P}=\overline{Q}-\overline{B}+\overline{A_B}=\overline{Q}-\overline{B}+\lambda\overline{B}\approx\overline{Q}-\overline{B}$. Through (53) , we have

$$\overline{B}=\overline{Q}-\overline{P}=\frac{1}{1-(1-\alpha)e^{-\theta}}. \quad (56)$$

Substituting (51) and (56) into (18), we obtain

$$\theta=nq(1-\alpha). \quad (57)$$

APPENDIX D
PROOF OF THEOREM 1 AND THEOREM 2

A. Proof of Theorem 1

Define $f(x)=\frac{x(e^x-1)}{e^x(e^x+x-1)}$. $f'(x)=\frac{(1-x)\left(e^{2x}-2e^x+1\right)+x^2}{e^x(e^x+x-1)^2}$. Since $e^x(e^x+x-1)^2>0$ is always true, here only the positive of $(1-x)\left(e^{2x}-2e^x+1\right)+x^2$ needs to be considered. Define $g(x)=(1-x)\left(e^{2x}-2e^x+1\right)+x^2$. It can be known by numerical calculation that there exists $x_0\approx 1.2515$ such that $g(x)>0$ when $0<x<x_0$ and $g(x)<0$ when $x>x_0$, i.e., x_0 is the maximum value point of function $f(x)$ at $(0,\infty)$. Thus, the maximum value of $f(x)$ at $(0,\infty)$ is $f(x_0)\approx 0.2384$. Using nq instead of x, the maximum access throughput λ^a_{max} in (31) and q^* in (32) can be obtained.

B. Proof of Theorem 2

Define $f(x)=\frac{x}{e^x+x-1}$. $f'(x)=\frac{(1-x)e^x-1}{(e^x+x-1)^2}$. Define $g(x)=(1-x)e^x-1$. Since $g'(x)=-xe^x<0$ when $x\in(0,\infty)$, $g(x)$ is monotonically decreasing function at $(0,\infty)$, i.e., $g(x)<g(0)=0$ is always true when $x\in(0,\infty)$. Thus, $f(x)$ is monotonically decreasing function at $(0,\infty)$ as well. When x approaches 0, its function value approaches $1/2$. Using nq instead of x, the maximum data throughput in (33) can be obtained.

REFERENCES

[1] H.-M. Wang, Q. Yang, Z. Ding, and H. V. Poor, "Secure short-packet communications for mission-critical IoT applications," *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 2565–2578, 2019.

[2] P. Castagno, V. Mancuso, M. Sereno, and M. A. Marsan, "A simple model of MTC flows applied to smart factories," *IEEE Trans. Mob. Comput.*, vol. 20, no. 10, pp. 2906–2923, 2020.

[3] W. Zhan and L. Dai, "Massive random access of machine-to-machine communications in lte networks: Modeling and throughput optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2771–2785, 2018.

[4] L. Dai and X. Sun, "A unified analysis of ieee 802.11 dcf networks: stability, throughput, and delay," *IEEE Trans. on Mobile Comput.*, vol. 12, no. 8, pp. 1558–1572, 2013.

[5] NR, *Medium Access Control (MAC) protocol specification (Release 17)*, document TS 38.321 V17.0.0, 3GPP, Mar. 2022.

[6] H. Zhou, Y. Deng, L. Feltrin, and A. Höglund, "Analyzing novel grant-based and grant-free access schemes for small data transmission," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2805–2819, 2022.

[7] X. Sun, H. Zhang, W. Zhan, X. Wang, and X. Chen, "How to survive 10 years' life time for machine type devices: a study of random access with sleeping-awake cycle," *IEEE Trans. Commun.*, vol. 71, no. 11, pp. 6727–6744, 2023.

[8] L. Kleinrock and S. S. Lam, "Packet-switching in a slotted satellite channel," in *Proceedings of the June 4-8, 1973, national computer conference and exposition*, 1973, pp. 703–710.

[9] N. Abramson, "The throughput of packet broadcasting channels," *IEEE Trans. Commun.*, vol. 25, no. 1, pp. 117–128, Jan. 1977.

[10] B. Hajek and T. Van Loon, "Decentralized dynamic control of a multiaccess broadcast channel," *IEEE Transactions on Automatic Control*, vol. 27, no. 3, pp. 559–569, Jun. 1982.

[11] L. Dai, "Stability and delay analysis of buffered Aloha networks," *IEEE Trans. Wireless Commun.*, 2012.

[12] Yang Yang and Tak-Shing Peter Yum, "Delay distributions of slotted Aloha and CSMA," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1846–1857, Nov. 2003.

[13] M. Rivero-Angeles, D. Lara-Rodriguez, and F. Cruz-Perez, "Differentiated backoff strategies for prioritized random access delay in multiservice cellular networks," *IEEE Trans. Veh. Technol.*, vol. 58, no. 1, pp. 381–397, Jan. 2009.

[14] 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.

[15] J.-B. Seo and V. C. M. Leung, "Design and analysis of backoff algorithms for random access channels in UMTS-LTE and IEEE 802.16 systems," *IEEE Trans. Veh. Technol.*, vol. 60, no. 8, pp. 3975–3989, Oct. 2011.

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

[17] H. Wu, C. Zhu, R. J. La, X. Liu, and Y. Zhang, "FASA: accelerated S-Aloha using access history for event-driven M2M communications," *IEEE/ACM Trans. Networking*, vol. 21, no. 6, pp. 1904–1917, Dec. 2013.

[18] C.-H. Wei, G. Bianchi, and R.-G. Cheng, "Modeling and analysis of random access channels with bursty arrivals in OFDMA wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 1940–1953, Apr. 2015.

[19] O. Arouk and A. Ksentini, "General model for RACH procedure performance analysis," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 372–375, Feb. 2016.

[20] 3rd Generation Partnership Project, "Ran improvements for machine-type communications," *V11.0.0*, Sep. 2011.

[21] Y.-J. Choi, S. Park, and S. Bahk, "Multichannel random access in OFDMA wireless networks," *IEEE J. Select. Areas Commun.*, vol. 24, no. 3, pp. 603–613, Mar. 2006.

[22] Y. Han, J. Deng, and Z. Haas, "Analyzing multi-channel medium access control schemes with Aloha reservation," *IEEE Trans. Wireless Commun.*, vol. 5, no. 8, pp. 2143–2152, Aug. 2006.

[23] A. Mutairi, S. Roy, and G. Hwang, "Delay analysis of OFDMA Aloha," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 89–99, Jan. 2013.

[24] J. Choi, "Fast retrial for low-latency connectivity in MTC with two different types of devices," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1786–1789, Oct. 2020.

[25] W. Crowther, R. Rettberg, D. Walden, S. Ornstein, and F. Heart, "A system for broadcast communication: reservation-Aloha," in *Proc. 6th Hawaii Int. Conf. Syst. Sci*, 1973, pp. 596–603.

[26] F. Tobagi and L. Kleinrock, "Packet switching in radio channels: Part III–Polling and (dynamic) split-channel reservation multiple access," *IEEE Trans. Commun.*, vol. 24, no. 8, pp. 832–845, Aug. 1976.

[27] Tasaka, "Stability and performance of the R-Aloha packet broadcast system," *IEEE Trans. Comput.*, vol. C-32, no. 8, pp. 717–726, Aug. 1983.

[28] S. Tasaka and Y. Ishibashi, "A reservation protocol for satellite packet communication–a performance analysis and stability considerations," *IEEE Trans. Commun.*, vol. 32, no. 8, pp. 920–927, Aug. 1984.

[29] J.-B. Seo and H. Jin, "S-Aloha systems with successive transmission: emulating CSMA system," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7544–7558, Nov. 2021.

[30] S. Ghez, S. Verdu, and S. Schwartz, "Stability properties of slotted Aloha with multipacket reception capability," *IEEE Trans. Automat. Contr.*, vol. 33, no. 7, pp. 640–649, Jul. 1988.

[31] Lang Tong, Qing Zhao, and G. Mergen, "Multipacket reception in random access wireless networks: from signal processing to optimal medium access control," *IEEE Commun. Mag.*, vol. 39, no. 11, pp. 108–112, Nov. 2001.

[32] Qing Zhao and Lang Tong, "A multiqueue service room MAC protocol for wireless networks with multipacket reception," *IEEE/ACM Trans. Networking*, vol. 11, no. 1, pp. 125–137, Feb. 2003.

[33] V. Naware, G. Mergen, and L. Tong, "Stability and delay of finite-user slotted Aloha with multipacket reception," *IEEE Trans. Inform. Theory*, vol. 51, no. 7, pp. 2636–2656, Jul. 2005.

[34] J. Liu, J.-B. Seo, and H. Jin, "Online transmission control for random access with multipacket reception and reservation," *IEEE Internet Things J.*, vol. 9, no. 22, pp. 23 160–23 175, Nov. 2022.

[35] E. Casini, R. De Gaudenzi, and O. Herrero, "Contention resolution diversity slotted Aloha (CRDSA): an enhanced random access scheme for satellite access packet networks," *IEEE Trans. Wireless Commun.*, vol. 6, no. 4, pp. 1408–1419, Apr. 2007.

[36] G. Liva, "Graph-based analysis and optimization of contention resolution diversity slotted Aloha," *IEEE Trans. Commun.*, vol. 59, no. 2, pp. 477–487, Feb. 2011.

[37] M. Lee, J.-K. Lee, J.-J. Lee, and J. Lim, "R-CRDSA: reservation-contention resolution diversity slotted Aloha for satellite networks," *IEEE Commun. Lett.*, vol. 16, no. 10, pp. 1576–1579, Oct. 2012.

[38] C. Stefanovic and P. Popovski, "Aloha random access that operates as a rateless code," *IEEE Trans. Commun.*, vol. 61, no. 11, pp. 4653–4662, Nov. 2013.

[39] E. Paolini, G. Liva, and M. Chiani, "Coded slotted Aloha: a graph-based method for uncoordinated multiple access," *IEEE Trans. Inform. Theory*, vol. 61, no. 12, pp. 6815–6832, Dec. 2015.

[40] J.-B. Seo, Y. Hu, and H. Jin, "Time-offset Aloha with SIC," *IEEE Trans. on Mobile Comput.*, pp. 1–13, 2023.

[41] Y. Gu, Y. Xu, B. Zhang, Y. Wang, and Z. Yang, "Towards the random multi-access in SIoT: a generalized deduplication based CRDSA mechanism," *IEEE Internet Things J.*, 2024.

[42] M. Shirvanimoghaddam, M. Condoluci, M. Dohler, and S. J. Johnson, "On the fundamental limits of random non-orthogonal multiple access in cellular massive iot," *IEEE J. Select. Areas Commun.*, vol. 35, no. 10, pp. 2238–2252, Oct. 2017.

[43] J. Choi, "NOMA-based random access with multichannel Aloha," *IEEE J. Select. Areas Commun.*, vol. 35, no. 12, pp. 2736–2743, Dec. 2017.

[44] L. Mai, Q. Zhang, and J. Qin, "System throughput maximization of uplink NOMA random access systems," *IEEE Commun. Lett.*, vol. 25, no. 11, pp. 3654–3658, Nov. 2021.

[45] H. Huang, T. Ye, T. T. Lee, and W. Sun, "Delay and stability analysis of connection-based slotted-Aloha," *IEEE/ACM Trans. Networking*, pp. 1–17, 2020.

[46] Nan Jiang and Yansha Deng and Xin Kang and Arumugam Nallanathan, "Random access analysis for massive IoT networks under a new spatio-temporal model: A stochastic geometry approach," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5788–5803, 2018.

[47] D. K. Kim, G. D. Georgiev, and N. V. Markovskaya, "A model of random multiple access in unlicensed spectrum systems," in *Proc. 2022 Wave Electron. Appl. Inf. Telecommun. Syst. (WECONF)*, 2022.

[48] J. F. Shortle and J. M. Thompson and D. Gross, *Fundamentals of Queueing Theory*. John Wiley & Sons, 2018.

[49] Istvan Mezo, *The Lambert W function: its generalizations and applications*. Boca Raton, FL: Chapman and Hall/CRC, 2022.

[50] Y. Li and L. Dai, "Maximum sum rate of slotted aloha with capture," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 690–705, Feb. 2016.

[51] M. Xie and J. Gong and X. Jia and X. Ma, "Age and Energy Tradeoff for Multicast Networks with Short Packet Transmissions," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 6106–6119, Sep. 2021.

[52] D. Bertsekas and J. N. Tsitsiklis, *Introduction to probability*. Athena Scientific, 2008.

Weilong Zhu (Student Member, IEEE) received the B.E. degree in communication engineering from the School of Electronics and Communication Engineering, Sun Yat-sen University (Shenzhen Campus), Shenzhen, China, in 2023. He is currently working toward the M.E. degree in information and communication engineering with the School of Electronics and Communication Engineering, Sun Yat-sen University. His research interests include random access, Internet of Things, and stochastic modeling of wireless network.

Wen Zhan (Member, IEEE) received the B.S. and M.S. degrees from the University of Electronic Science and Technology of China, China, in 2012 and 2015, respectively, and the Ph.D. degree from the City University of Hong Kong, China, in 2019. He was a Research Assistant and a Post-Doctoral Fellow with the City University of Hong Kong. Since 2020, he has been with the School of Electronics and Communication Engineering, Sun Yat-sen University, China, where he is currently an Associate Professor. His research interests include Internet of things, modeling and performance optimization of next-generation mobile communication systems, reinforcement learning, queueing theory and its application in wireless communications.

Xinghua Sun (M'13) received the B.S. degree from Nanjing University of Posts and Telecommunications (NJUPT), China, in 2008 and the Ph.D. degree from the City University of Hong Kong (CityU), China, in 2013. In 2010, he was a visiting student with the National Institute for Research in Digital Science and Technology (INRIA), France. In 2013, he was a postdoctoral fellow at CityU. From 2015 to 2016, he was a postdoctoral fellow at the University of British Columbia, Canada. From July to Aug. 2019, he was a visiting scholar at Singapore University of Technology and Design, Singapore. From 2014 to 2018, he was an associate professor with NJUPT. Since 2018, he has been an associate professor with Sun Yat-sen University, Guangdong, China. Dr. Sun was a co-recipient of the Best Paper Award from the EAI IoTaaS in 2023 and the IEEE FCN in 2024. He served as the Technical Program Committee Member and the Organizing Committee Member for numerous conferences. His research interests are in the area of stochastic modeling of wireless networks and machine learning for next generation wireless communications and networks.

Xiang Chen Xiang Chen (Member, IEEE) received the B.E. and Ph.D. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2002 and 2008, respectively. From July 2008 to December 2014, he was with the Wireless and Mobile Communication Technology Research and Development Center (Wireless Center) and the Aerospace Center, Tsinghua University. In July 2005 and from September 2006 to April 2007, he visited NTT DoCoMo Research and Development (YRP), and Wireless Communications and Signal Processing (WCSP) Laboratory, National Tsing Hua University. Since January 2015, he has been with the School of Electronics and Information Technology, Sun Yat-sen University, where he is currently a Full Professor. His research interests mainly focus on 5G/6G wireless communications, big data and the Internet of Things (IoT).

Yuan Jiang (Member, IEEE) received the B.Eng. degree in satellite communication and the M.Sc. degree in communication and electronic system from the Information Engineering University, Zhengzhou, China, in 1991 and 1998, respectively, and the Ph.D. degree in information and communication engineering from Zhejiang University, Hangzhou, China, in 2004. From 1991 to 2005, he was with the Department of Communication Engineering, Information Engineering University. From 2005 to 2008, he was with the Postdoctoral Workstation of Computer Science and Engineering, South China University of Technology, Guangzhou, China. From 2008 to 2018, he was the Vice General Manager, a Chief Engineer, and the Vice President in a listed company. Since 2018, he has been a Professor, a Doctoral Supervisor, and the Director of the Provincial Key Laboratory, Sun Yat-sen University. He has presided or participated in more than 20 research projects, including the National Key Research and Development Program Special Project. His research interests include cognitive communications, satellite communication networks, satellite navigation, and their applications to wireless communication systems.