Statistical Delay Analysis of TDMA and ALOHA in Wireless Multihop Networks

Min Xie and Martin Haenggi

Department of Electrical Engineering, University of Notre Dame

Notre Dame, IN 46556, USA

{mxie—mhaenggi}@nd.edu

Abstract

We employ discrete-time queueing theory to analyze wireless multihop networks and derive an explicit end-to-end (e2e) delay distribution. The analysis takes into account the scheduling scheme, which is a two-level problem including both the (global) channel access scheduler (MAC) and the (local) packet scheduler at each node. Two MAC schemes, TDMA and ALOHA, are considered. Queueing models are established in such a way that the access delay can be incorporated into the service process. Characterizing the wireless channel by a capture model, the individual nodes are modeled as GI/Geom/1 queueing systems. However, in TDMA with constant bit rate (CBR) traffic, the corresponding D/Geom/1 system has a non-integer interarrival time and thus cannot be analyzed as usual. We use an eigenvalue approach to derive closed-form queue length and delay distributions of such D/Geom/1 systems. The error-prone wireless channel transforms the smooth and deterministic CBR traffic to bursty and correlated on-off, causing correlations in the delays at different nodes. We propose an approximative approach that includes the long-distance correlations in the correlation between the neighboring nodes. Based on a set of simulation results, an empirical model for the correlation coefficient is established. Then the network is modeled as a series of independent servers, but the delay variances are scaled by the correlation coefficient. Finally, TDMA and ALOHA are quantitatively compared in terms of the delay and the delay outage probability.

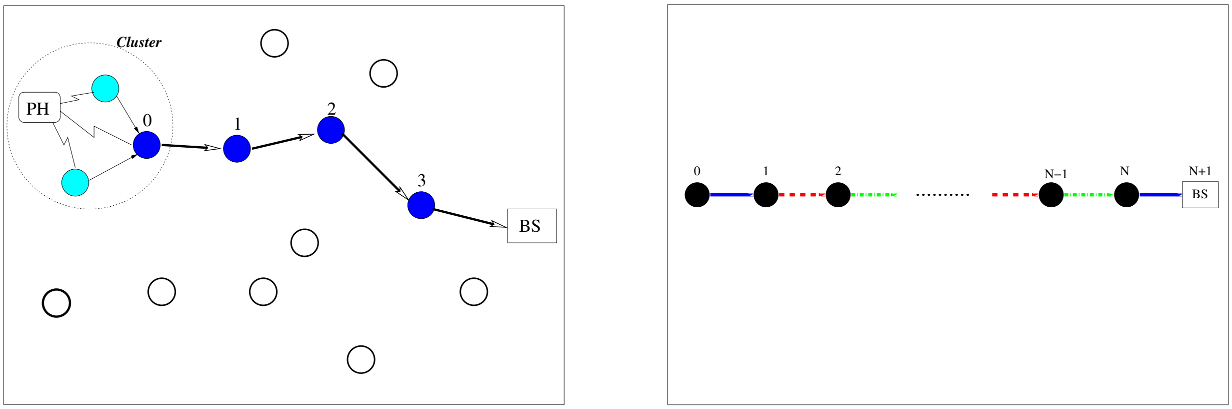
The support of the Center for Applied Mathematics (CAM) Fellowship of the University of Notre Dame and the partial support of NSF (grants ECS03-29766 and CAREER CNS04-47869) are gratefully acknowledged.

I. Introduction

With the growing demand for real-time applications over wireless networks, increasing attention is paid to the delay analysis over error-prone channels. In multihop networks, like ad hoc, mesh, and multihop cellular networks, the analysis is more challenging due to the delay accumulation at each hop. Many factors interact with each other, including the routing algorithm, the scheduling algorithm, the wireless channel and the resulting interference (*e.g.*, in Fig. 1(a), two types of interference exist, *inter-flow interference* caused by cross-traffic flows and *intra-flow interference* caused by multiple nodes of a single flow [1]). The analysis is unlikely to be tractable if all these factors are considered together.

Our focus is on scheduling, which is a two-level problem, including the global MAC scheme and the local packet scheduling algorithm. We assume a single flow so that the routing algorithm and the inter-flow interference can be neglected. This is a common scenario, *e.g.*, in sensor networks with a single phenomenon to be detected. In Fig. 1(a), a set of nodes around the phenomenon periodically detect, collect, and then forward the sensed data to a cluster head (node 0). A path is established from the cluster head to the base station (BS). Then, the two-dimensional (2-D) topology is reduced to one-dimensional (1-D), which can be further simplified to a regular line network (Fig. 1(b)). Due to the zero inter-flow interference assumption, the analysis of the regular line network provides an upper performance bound for general 2-D networks.

For a single flow case, FIFO is sufficient for local scheduling, but the MAC scheme needs to be carefully designed. Wireless MAC schemes are designed to combat interference and take advantage of spatial reuse to enhance the throughput. Doing so, extra access delays may be accumulated. It is intriguing to study how the improved throughput affects the delay. Since the delay analysis involves queueing theory, it might be intractable if the MAC scheme is too complicated. For a tractable analysis, we consider two simple but typical MAC schemes, m-phase spatial TDMA [2] and slotted ALOHA. In TDMA, a node is scheduled to transmit once in m time slots, and nodes m hops apart can transmit simultaneously. In ALOHA, every node



(a) General network (PH: phenomenon)

(b) Line network

Fig. 1. Wireless multihop networks

independently transmits with probability p_m whenever it has packets. TDMA (with nodes fully cooperative) and ALOHA (with nodes completely independent) represent two extremes in terms of the level of the node coordination and thus provide an upper and lower performance bound for other meaningful MAC schemes.

A. Previous work

The throughput and single-hop delay of many MAC schemes have been comprehensively studied in the literature [2], [3]. However, little work has been done on their multihop delay. Moreover, the study of MAC schemes usually focuses on the access delay without considering local scheduling by assuming that traffic is generated in a way that there is no queueing delay. For example, in an "infinite population model", a new node is generated to represent the newly generated packet; or new packets are generated only when the buffer is empty [3]–[7]. These models are simplified and unrealistic. In practice, the number of nodes is finite, and new packets may be generated when the buffer is non-empty and then be queued. Local scheduling is required to determine the transmission order of queued packets.

The impact of the wireless channel is another important MAC-related issue [3]–[5], [7]. With the "capture" property of the radio receiver, the throughput of slotted ALOHA is improved [3], [4], [8]. In some MAC schemes, using the channel characteristics to control the backoff timer or the transmit power, the delay and throughput can be changed [6], [7]. Therefore, the channel model is critical for the analysis. The "disk model" [9], assuming a fixed transmission range, has been very popular because of its simplicity. In m-phase TDMA with m chosen appropriately (*e.g.*, $m = 3$ in [10]), the disk model results in error-free channels since the simultaneous transmissions of nodes m hops apart cause no interference or collisions [10], [11]. In reality, the transmission success depends on the received signal-to-noise-and-interference ratio (SNIR). So TDMA does not completely eliminate the interference unless spatial reuse is completely foregone. The "capture model" [12] is more appropriate, in which packet transmissions fail if the received SNIR is too low. The failed packets have to wait for the next transmission opportunity. The packets behind them need a local scheduler to solve the competition for the transmission opportunities.

With practical traffic and channel models, combining scheduling with MAC schemes is non-trivial. However, from the perspective of queueing theory, the joint queueing analysis might be difficult. For instance, the queueing analysis of multihop networks usually neglects the MAC-dependent access delay [10], [13]–[15]. Closed-form solutions for the delay of the regular line network with a single source (like Fig. 1(b)) are provided when the traffic is geometric [14] or the channel is error-free [15]. For other traffic models and error-prone wireless channels, some approximations are needed, *e.g.*, the "independence" assumption: In [16], the delay variance of a two-node tandem network is derived by assuming that the two nodes are independent. Similarly, in [17], an IEEE 802.11 wireless ad hoc network is modeled as a series of *independent* M/G/1 systems to obtain a product-form delay distribution. However, in most cases the "independence" assumption does not hold and would lead to a very loose performance expression.

B. Our contributions

This paper aims to derive explicit e2e delay distribution of a wireless regular line network (Fig. 1(b)) with its two-level scheduling problem. Local packet scheduling causes a queueing delay, while the global MAC scheme results in an access delay. Our contributions are two-fold. First, we use discrete-time queueing theory to analyze the delay of the individual nodes, accounting for both the queueing delay and the access delay. The channel is characterized by a capture model. Traffic is CBR, which covers many real-time applications, *e.g.*, voice data [18] and periodic traffic in sensor networks. Unlike previous work, the traffic generation is not controlled by the node buffer and the network stability is guaranteed, *i.e.*, the nodes are not always busy. The queueing model is established in such a way that the MAC-dependent access delay can be directly incorporated into the service process. Then, the nodes can be modeled as GI/Geom/1 systems.

With CBR traffic, the source node is modeled as D/Geom/1. In TDMA, in order to establish a D/Geom/1 model, the interarrival time becomes a non-integer constant, a case that has not been considered in conventional discrete-time queueing theory. We establish a 2-D Markov chain and employ an eigenvalue approach to derive a closed-form queue length and delay distribution of such D/Geom/1 systems, thereby generalizing the analysis of the conventional D/Geom/1 system with integer interarrival time. Then we characterize the output by an on-off process. For very heavy traffic, where the interarrival time and the intertransmission time (frame length m) differs only by one time slot, a 1-D delay model is used for the delay analysis.

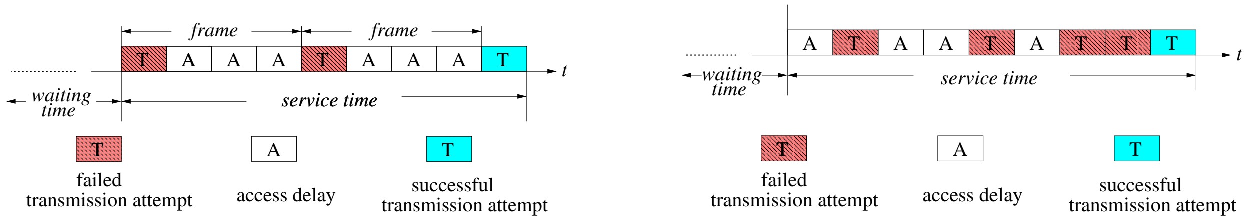
For CBR traffic, correlations exist between the delays experienced at different nodes, indicating that the "independence" assumption does not hold. Hence our second contribution is an approximative approach to simplify the correlation analysis. Based on the simulation results, an empirical model for the correlation coefficient is established. Then, the line network is treated like a series of independent GI/Geom/1 systems, but the delay variances are scaled by this correlation coefficient. With the complete e2e delay distribution, we calculate the delay outage probability (the probability that the e2e delay exceeds a predetermined threshold) for delay-sensitive applications. As a complement to previous work focusing on the throughput, this paper compares TDMA and ALOHA in terms of the e2e delay and the outage probability.

The rest of the paper is organized as follows. The approximative approach is introduced in Section II. Then, the TDMA and ALOHA networks are analyzed individually in Sections III and IV, including the derivation of the delays at the individual nodes and the correlation coefficient. In Section V, TDMA and ALOHA are compared in terms of the e2e delay and delay outage probability. Section VI concludes the paper.

II. System Model

The regular line network under consideration (Fig. 1(b)) is composed of $N+1$ transmitting nodes. Denote node i by n_i and the delay experienced at n_i by W_i. More precisely, W_i is the interval between the instant that the packet is successfully received by n_i and the instant that the packet is successfully transmitted and thus received by n_{i+1}. The e2e delay is the sum of the W_i's. A FIFO discipline is used at the nodes. A CBR flow of fixed-length packets is generated at the source n_0, and all remaining nodes except the BS are pure relays. The time is slotted to one packet transmission duration, and the network is modeled as a discrete-time tandem queueing network. The traffic interarrival time is r slots ($r \in \mathbb{N}$). The channels are assumed to be subject to independent errors with capture probability p_s (*e.g.*, AWGN or block fading channels). The network is 100% reliable, *i.e.*, the failed packets will be retransmitted at each link until received successfully.

In queueing theory, the delay W_i is composed of two parts, the waiting time (from the arriving instant to the instant that the packet is about to be served) and the service time, as shown in Fig. 2. In TDMA, the node is given a transmission opportunity once in m time slots. When dividing the time into frames of m slots and setting the beginning of a frame as the slot allocated to the node, the service time is geometric with p_s, denoted by $\mathcal{G}_{p_s}$. The access delay is hidden in the frame so that each node can be modeled as a GI/Geom/1 system at the frame level. In ALOHA, busy nodes make a transmission decision at every time slot. A packet is correctly received if and only if the node attempts to transmit and the transmission is successful, with probability $s \triangleq p_s p_m$ (given that the arrival and the channel state are independent[1]). Overall, the service time is $\mathcal{G}_s$. This way, it is not necessary to distinguish the access delay from the failed transmission attempts. Then each node can be modeled as a GI/Geom/1 at the slot level.


frame
frame
T
A
A
A
T
A
A
A
T
t
waiting time
service time
T
failed transmission attempt
A
access delay
T
successful transmission attempt
A
T
A
A
T
A
T
T
T
t
waiting time
service time
T
failed transmission attempt
A
access delay
T
successful transmission attempt

(a) TDMA

(b) ALOHA

Fig. 2. Packet transmission procedure in TDMA and ALOHA

The individual nodes can be analyzed based on the established GI/Geom/1 models, deriving the complete distribution of W_i, with the mean denoted by μ_i and the variance by σ_i^2. The e2e analysis requires the output characterization of each node. It is known that except for memoryless Poisson and geometric traffic, the output of node n_i, or the input to the following node n_{i+1}, depends on the input and service process of n_i [14]. Therefore, the delay W_{i+1} is correlated with the delay W_i. In a tandem network, simulation results reveal that the delay W_{i+1} is correlated with all W_j for $j \leq i$.

It is hardly feasible to explicitly derive the e2e delay distribution if all correlations are considered. Instead, we propose an approximative approach that includes the long-distance correlations in the correlation between neighboring nodes. Then the line network can be modeled as a series of queueing systems which are correlated only with their neighbors. Since all nodes have an identical channel and all nodes except for the source node are pure relays, we further

[1] To account for the half-duplex restriction, here p_s is the conditional capture probability given that the receiver is listening.

assume that the correlation coefficient $\eta \triangleq \text{cor}(W_i, W_{i+1})$ is independent of i. Based on the analysis of W_i, the e2e delay mean and variance are:

$$\mu = \sum_{i=0}^{N} \mu_i, \quad \sigma^2 = \sum_{i=0}^{N} \sigma_i^2 + \sum_{i=0}^{N-1} \eta \sigma_i \sigma_{i+1}. \tag{1}$$

Since the correlation does not affect the mean, the e2e delay mean is simply the sum of the μ_i's (Fig. 3(a)). For the variance, if the W_i's were independent ($\eta = 0$), then σ^2 would be the sum of σ_i^2, like in previous work [16], [17]. However, the W_i's are not independent ($\eta \neq 0$), and a gap appears between the real e2e delay variance (the "simulation" line in Fig. 3(b)) and the sum of σ_i^2 (the "independence" line). The variance σ^2 is either greater or smaller than the sum of σ_i^2, depending on whether the correlation is positive or negative.

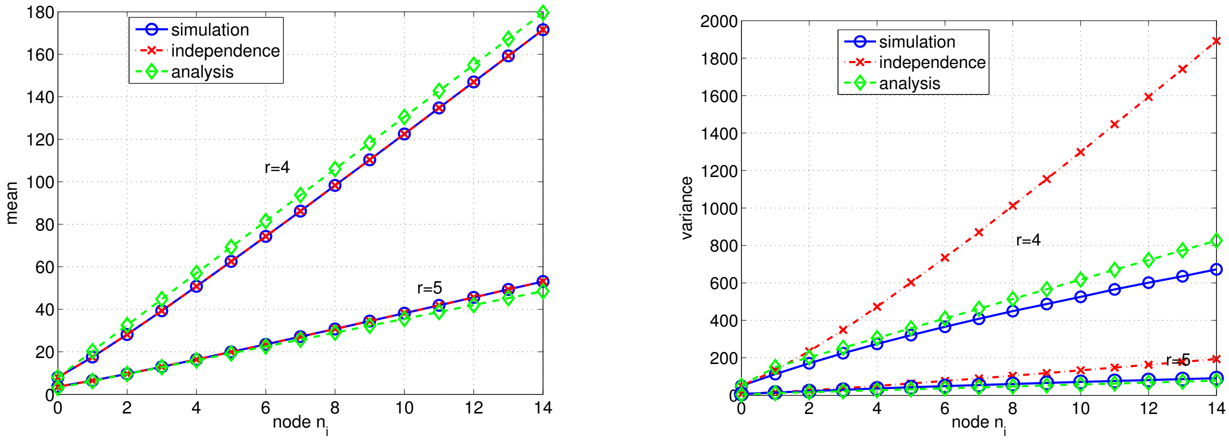
Intuitively, for bursty traffic, if the packet experiences a long delay at n_i because of a bad channel or the long accumulated queue, node n_{i+1} probably has cleared its buffer during this period so that when the packet arrives at n_{i+1}, it is likely to experience a short delay, and vice versa. In other words, the correlation is expected to be negative ($\eta < 0$), hence the e2e delay variance σ^2 is smaller than the sum of σ_i^2. This effect becomes more prominent as the traffic intensity increases. So, it is reasonable to conclude that the correlation is a function of the intensity. The detailed derivation of the correlation coefficient η will be given in the following sections for TDMA and ALOHA separately.

III. Analysis of the m-Phase TDMA Network

In m-phase TDMA networks, each node can be modeled as a GI/Geom/1 system with service rate p_s at the frame level. The average arrival rate is m/r (m/r is a reduced fraction). We further assume $r < 2m$ for heavy traffic. Define the traffic intensity as $\rho \triangleq m/(rp_s)$ and constrain it to $\rho < 1$ for stability. Note that previous MAC-related work implicitly assumed $\rho = 1$.

A. Delay analysis of the source node n_0

The source n_0 is fed with a CBR flow of interarrival time r. At the frame level, n_0 is modeled as a D/Geom/1 but the interarrivel time r/m is an integer for $m < r < 2m$. The number of



(a) Delay mean

(b) Delay variance

Fig. 3. Comparison of the e2e delay performance in the TDMA network with $m = 3, p_s = 0.8, N = 14, r = 4, 5$

arrivals in one frame jumps between 0 and 1, depending on the arrival patterns of all previous $r - 1$ frames. Such a D/Geom/1 system has not been studied by conventional queueing theory, where it is assumed that the interarrival times are all integer and independent.

We establish a 2-D Markov chain $(L(t), U(t))$ to characterize the system state at the beginning of frame t, where $L(t) \in \mathbb{N}$ is the queue length, and $U(t) \in \{1, \ldots, r\}$ is the number of *time slots* till the next packet arrival. All state transitions occur at the frame boundaries. Define $\Delta \triangleq r - m$. If $1 \le U(t) \le m$, a packet will arrive during frame t. At frame $t + 1$, $U(t)$ will evolve to $U(t+1) = U(t) + r - m = U(t) + \Delta > \Delta$. If $m < U(t) \le r$, no packet arrives during frame t, and $U(t+1) = U(t) - m \le r - m = \Delta$. Divide the state space $\{1, \ldots, r\}$ of $U(t+1)$ into two subspaces, $\mathbb{U}_0 = \{1, 2, \ldots, \Delta\}$ and $\mathbb{U}_1 = \{\Delta + 1, \Delta + 2, \ldots, r\}$, where the subscripts represent the number of packets arriving prior to the beginning of frame $t + 1$ (or during frame t). The equilibrium state probability is defined as $Q_l(u) \triangleq \lim_{t\to\infty} \Pr\{L(t) = l, U(t) = u\}$, which is calculated in Theorem 3.1.

Theorem 3.1: Consider a D/Geom/1 system with interarrival time r/m ($r, m \in \mathbb{N}$ and r/m is a reduce fraction) and service rate p_s. The equilibrium probability $Q_l(u)$ for $l \geq 1$ is

$$Q_l(u) \approx \rho \lambda_0^l a^{u-1}(a-1), \quad \text{where} \quad a = \lambda_0^{-1/r}, \tag{2}$$

and λ_0 is a real positive eigenvalue of the matrix $\mathbf{M}(\lambda)$

$$\mathbf{M}(\lambda) = \mathbf{M_0} + \mathbf{M_1}\lambda + \mathbf{M_2}\lambda^2, \tag{3}$$

with

$$\mathbf{M_0} = \begin{bmatrix} \mathbf{0} & (1-p_s)\mathbf{I}_m \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathbf{M_1} = \begin{bmatrix} \mathbf{0} & p_s\mathbf{I}_m \\ (1-p_s)\mathbf{I}_\Delta & \mathbf{0} \end{bmatrix} - \mathbf{I}, \quad \mathbf{M_2} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ p_s\mathbf{I}_\Delta & \mathbf{0} \end{bmatrix}. \tag{4}$$

Proof: Following the evolution of the system states, we have the state transitions

$$\begin{aligned}
\text{For} \quad u \in \mathbb{U}_0: \quad Q_l(u) &= \begin{cases} (1-p_s)Q_l(u+m) + p_s Q_{l+1}(u+m) & l > 0 \\ p_s Q_1(u+m) & l = 0 \end{cases} \\
\text{For} \quad u \in \mathbb{U}_1: \quad Q_l(u) &= \begin{cases} (1-p_s)Q_{l-1}(u-\Delta) + p_s Q_l(u-\Delta) & l > 1 \\ Q_0(u-\Delta) + p_s Q_1(u-\Delta) & l = 1. \end{cases}
\end{aligned} \tag{5}$$

Note that $Q_0(u) = 0$ for all $u \in \mathbb{U}_1$. Define $\mathbf{v}_l := \{Q_l(1), Q_l(2), \ldots, Q_l(r)\}$ as a state vector. Since it is a probability distribution, $\sum_{l=0}^{\infty}(\mathbf{v}_l \cdot \mathbf{e}) = 1$, where $\mathbf{e}$ is a column vector with all its entries equal to 1. For $l \geq 1$, rewrite (5) in matrix form,

$$\mathbf{v}_l\mathbf{M}_0 + \mathbf{v}_{l+1}\mathbf{M}_1 + \mathbf{v}_{l+2}\mathbf{M}_2 = 0, \tag{6}$$

which can be solved by an eigenvalue approach [19]. First solve $\boldsymbol{\phi}(\lambda)\mathbf{M}(\lambda) = 0$, where $\mathbf{M}(\lambda)$ is given in (3) and $\boldsymbol{\phi}(\lambda) = \{\phi_u(\lambda) | u = 1, \ldots, r\}$ and λ are the left eigenvector and eigenvalue of $\mathbf{M}(\lambda)$. Second, given $C = \text{degree}\{\det[\mathbf{M}(\lambda)]\}$ distinct eigenvalues $\{\lambda_i \mid i = 0, 1, \ldots, C-1\}$,

$$\mathbf{v}_l = \sum_{i=0}^{C-1} c_i \boldsymbol{\phi}(\lambda_i)\lambda_i^l, \quad l \geq 1, \tag{7}$$

where c_i is any constant. To form a probability distribution, only the eigenvalues inside the unit circle ($|\lambda_i| < 1$) are included. Therefore, set $c_i = 0$ for $|\lambda_i| \geq 1$. In view of the matrices $\mathbf{M}_0$,

$\mathbf{M}_1$, and $\mathbf{M}_2$, the expansion of $\boldsymbol{\phi}(\lambda)\mathbf{Q}(\lambda) = 0$ gives

$$\phi_u(\lambda) = \begin{cases} \beta\phi_{u+m} & u \le \Delta \\ \beta\lambda^{-1}\phi_{u-\Delta} & u > \Delta. \end{cases} \quad \text{where} \quad \beta = 1 - p_s + p_s\lambda. \tag{8}$$

Multiplying $\phi_1(\lambda)\phi_2(\lambda)\cdots\phi_r(\lambda)$ in (8) yields

$$\lambda = (1 - p_s + p_s\lambda)^{r/m} = \beta^{r/m}. \tag{9}$$

Set $x = \lambda^{1/r}$ and rewrite (9) in a polynomial form,

$$p_s x^r - x^m + 1 - p_s = 0, \tag{10}$$

with roots $\{x_i\}$ corresponding to the eigenvalues as in $\lambda_i = x_i^r$. As pointed out in [20] and Descartes' Sign Rule, the polynomial (10) has m roots inside the unit circle, among which only one is real positive, denoted by x_0, and all others are complex or negative. With the involvement of complex and negative eigenvalues, $\mathbf{v}_l$ becomes complicated. Since the real positive root x_0 is the largest among these m roots[2], we simplify $\mathbf{v}_l$ to

$$\mathbf{v}_l \approx c_0 \boldsymbol{\phi}(\lambda_0)\lambda_0^l. \tag{11}$$

The simplification gets tight when $p_s \to 1$ and λ_0 is much larger than other eigenvalues. To solve the corresponding eigenvector $\boldsymbol{\phi}(\lambda_0)$, exploring (8) gives

$$\phi_{u+1}(\lambda_0) = a\phi_u(\lambda_0) \quad \Longrightarrow \quad \phi_u(\lambda_0) = a^{u-1}\phi_1(\lambda_0), \quad u \ge 1, \tag{12}$$

where

$$a = \begin{cases} \beta_0^{a_1}\lambda_0^{a_2-a_1}, & \text{if} \quad \Delta = (a_2 r + 1)/a_1 \\ \beta_0^{-a_1}\lambda_0^{a_1-a_2}, & \text{if} \quad \Delta = (a_2 r - 1)/a_1. \end{cases} \quad a_1, a_2 \in \mathbb{N}, \quad \beta_0 = 1 - p_s + p_s\lambda_0. \tag{13}$$

Combining (9) and (13) leads to several equations on a,

$$a = \lambda_0^{-1/r} = \frac{1}{x_0}, \quad a^\Delta = \frac{\beta_0}{\lambda_0}, \quad a^m = \frac{1}{\beta_0}. \tag{14}$$

[2] This is confirmed by extensive numerical results. Also, in Proposition 3.3, x_0 is shown to be very close to 1.

Based on (11), (12), and (14), the equilibrium state probability $Q_l(u)$ is

$$Q_l(u) \approx \rho(1-\lambda_0)\lambda_0^{l-1}a^{u-1}\frac{1-a}{1-a^r}, \tag{15}$$

which is further simplified to (2). ■

To validate the analysis, consider $m = 1$ (which is not practical for wireless TDMA due to the half-duplex restriction), when the interarrival time is an integer r and the system reduces to a conventional D/Geom/1 system Note that $m = 1$. According to Theorem 3.1, there is exactly one eigenvalue inside the unit circle. Then, the calculation of $Q_l(u)$ is exact. Plugging $m = 1$ into (2), we deduce the queue length probability $\pi_l := \sum_{u=1}^{r} Q_l(u)$ as follows

$$\pi_l = \begin{cases} \rho(1-\lambda_0)\lambda_0^{l-1} & l > 0 \\ 1-\rho & l = 0, \end{cases} \tag{16}$$

where λ_0 is a solution to $\lambda = (p_s\lambda + 1 - p_s)^r$. This is consistent with the well-known results for conventional D/Geom/1 queues [21], [22]. Therefore, Theorem 3.1 provides a generalized queueing analysis of D/Geom/1 systems even if the interarrival time is not integer. With $Q_l(u)$, the generalized delay distribution of D/Geom/1 systems is given in the following theorem.

Theorem 3.2: Consider a general D/Geom/1 queueing system with interarrival time r/m and service rate p_s. Then, the probability generating function (pgf) of the delay distribution is

$$G_{W_0}(z) \approx \frac{(a-1)(a^m - z^m)\beta_0 z^m}{(a-z)(1-\beta_0 z^m)}, \tag{17}$$

where a and β_0 are given in (13).

Proof: Divide the time unit into m sub-time units and use these sub-time units to measure the delay. The delay is composed of three independent parts, the access delay $u_a \in \{0, \ldots, m-1\}$, the queueing delay S_Q, and the service time S. At the packet arriving instant, an access delay u_a implies $U(t) = r - u_a$. The probability that the newly arriving packet with an access delay u_a sees $l-1$ packets in the system is

$$Q_l^*(u_a) = \frac{Q_l(r-u_a)}{\sum_{l=1}^{\infty}\sum_{u\in\mathbb{U}_1} Q_l(u)} = \frac{r}{m}Q_l(r-u_a) \approx \frac{\lambda_0^l}{p_s}a^{r-1-u_a}(a-1), \tag{18}$$

where $\sum_{l=1}^{\infty} \sum_{u \in \mathbb{U}_1} Q_l(u) = m/r$ is the probability that an arrival is about to occur [22]. The packet service time is geometric as follows (where $d_{km} \triangleq \Pr\{S = km\}$)

$$\begin{aligned} d_{km} &= p_s(1-p_s)^{k-1}, \quad k \geq 1 \quad \Longrightarrow \\ G_S(z) &= \sum_{k=1}^{\infty} p_s(1-p_s)^{k-1} z^{km} = \frac{p_s z^m}{1-(1-p_s)z^m}. \end{aligned} \tag{19}$$

S_Q is the sum of $l-1$ such independent service times. So the pgf is $G_{S_Q}(z) = [G_S(z)]^{l-1}$. The total delay $W_0 = u_a + S_Q + S$ has a pgf

$$G_{W_0}(z) = \sum_{l=1}^{\infty} \sum_{u_a=0}^{m-1} Q_l^*(u_a) G_{S_Q}(z) G_S(z) z^{u_a}, \tag{20}$$

which is simplified to (17) using the approximation in (18). ■

Note that in TDMA, because of the shift of the frame beginnings, the pmf and pgf of the service time of the newly arriving packet must be modified to

$$\begin{aligned} d_{km+1} &= p_s(1-p_s)^{k}, \quad k \geq 0 \quad \Longrightarrow \\ G_S(z) &= \sum_{k=0}^{\infty} p_s(1-p_s)^{k} z^{km+1} = \frac{p_s z}{1-(1-p_s)z^m}. \end{aligned} \tag{21}$$

The pgf of the delay W_0 is thus

$$G_{W_0}(z) \approx \frac{(a-1)(a^m - z^m)\beta_0 z}{(a-z)(1-\beta_0 z)}, \quad a = 1/x_0, \quad \beta_0 = a^m. \tag{22}$$

This pgf is determined by x_0, the real positive root of the polynomial (10) inside the unit circle. Usually x_0 has only numerical solutions for large m and r. We derive a closed-form expression for x_0 in Proposition 3.3.

Proposition 3.3: Consider the polynomial $p_s x^r - x^m + 1 - p_s = 0$. The real positive root x_0 inside the unit circle can be well approximated by

$$x_0 \approx 1 - \frac{2(1-\rho)}{\Delta\rho}, \quad \rho = m/(rp_s), \quad \Delta = r - m < m. \tag{23}$$

Proof: Based on Descartes' Sign Rule, there are exactly two real positive roots, one of which is 1 and the other is $x_0 \in (0,1)$. A single local minimum $x_{min} = \rho^{\frac{1}{\Delta}} < 1$ lies between

x_0 and 1. Using two inequalities [23]

$$-\frac{1-\rho}{\rho} < \ln\rho \leq \Delta(\rho^{\frac{1}{\Delta}} - 1), \quad \rho < 1 \text{ and } \Delta \in \mathbb{N}, \tag{24}$$

x_{min} is lower bounded as $x_{min} \gtrapprox 1 - (1-\rho)/(\Delta\rho)$. Assume an equal distance from x_{min} to these two roots, *i.e.*, $1 - x_{min} \approx x_{min} - x_0 \Longrightarrow x_0 \approx 2x_{min} - 1$. Then x_0 is solved as in (23). ■

The approximation (23) is tight when Δ is large and/or ρ is close to 1, both of which also guarantee $x_0 \lessapprox 1$. The delay mean and variance can be directly calculated from $G_{W_0}(z)$ (22),

$$\begin{aligned} \mu_0 &\approx \frac{a}{a-1} \approx \frac{\Delta\rho}{2(1-\rho)}, & (25)\\ \sigma_0^2 &\approx \frac{a(2-a)}{(a-1)^2} \approx \Big(\frac{\Delta\rho}{2(1-\rho)}\Big)^2 - \frac{\Delta\rho}{1-\rho} = \mu_0(\mu_0 - 2). & (26)\end{aligned}$$

Since the approximation (23) is more accurate for large Δ, the special case $\Delta = 1$ will be studied separately. Instead of the 2-D Markov chain $\{L(t), U(t)\}$, we use a 1-D delay model in Theorem 3.4 to directly track the delay evolution of the Head of Line (HOL) packet.

Theorem 3.4: Consider a D/Geom/1 queueing system with interarrival time $(m+1)/m$ and service rate p_s. Then, the pgf and the mean and variance of the delay W_0 are

$$\begin{aligned} G_{W_0}(z) &= \frac{(z^m - 1)z}{z(1-z^m) - p_s(1-z^{m+1})} \cdot \frac{1-\rho}{\rho} & (27)\\ \mu_0 &= \frac{1}{2(1-\rho)}, \qquad \rho = \frac{m}{rp_s} & (28)\\ \sigma_0^2 &= \frac{1}{4(1-\rho)^2} - \frac{m+2}{6(1-\rho)}. & (29)\end{aligned}$$

Proof: Using the delay model [24], the system state is denoted by the delay of the HOL packet in terms of time slots. But state transitions occur at the frame boundaries. The transition probabilities Z_{jk} are

$$Z_{jk} = \begin{cases} p_s & j \geq 0, \quad k = j-1, \\ 1-p_s & j \geq 0, \quad k = j+m, \\ 1 & j = -1, \quad k = m-1, \end{cases} \tag{30}$$

The negative state -1 indicates that the buffer is empty and the remaining time prior to the next arrival is one slot. Let the steady-state probabilities be $\{z_j | j \geq -1\}$ and the delay probability be $\{\pi_j | j \geq 1\}$. Then,

$$\pi_j = \frac{z_{j-1}p_s}{\sum\limits_{k=0}^{\infty} z_k p_s} = \frac{z_{j-1}}{\sum\limits_{k=0}^{\infty} z_k}, \tag{31}$$

where $\sum_{k=0}^{\infty} z_k p_s$ is the normalization constant. The factor p_s is needed since the erroneous packets are not included to calculate the delay distribution. (30) gives a set of balance equations

$$\pi_j = \begin{cases} p_s\pi_{j+1} & 1 \leq j < m \\ p_s\pi_{m+1} + p_s\pi_1 & j = m \\ p_s\pi_{j+1} + (1-p_s)\pi_{j-m} & j > m, \end{cases} \tag{32}$$

which leads to the pgf $G_{W_0}(z)$ with an unknown parameter π_1. Based on $G_{W_0}(1) = 1$, π_1 is deduced to $\pi_1 = (1-\rho)/(p_s\rho)$. Note that the number of unknown parameters involved is Δ. That is why this approach is applied to $\Delta = 1$ only. The mean (28), variance (29), and the pmf of W_0 can be derived from the pgf (27) in a straightforward manner. ■

Comparing the approximate delay analysis (25), (26) (setting $\Delta = 1$) and the accurate analysis (28), (29), we find that even for small Δ, the approximation is tight if $\rho \to 1$.

B. Delay analysis of the relay nodes

The relay nodes are fed with the output of the source node n_0. Since the packet departure occurs only at the frame boundaries, the interarrival time is integer for relay nodes n_i ($1 \leq i \leq N$) that are modeled as GI/Geom/1 and analyzed conventionally. We first analyze their input, or the output of n_0 in Theorem 3.5.

Theorem 3.5: Consider a D/Geom/1 queueing system with service rate p_s and interarrival time r/m ($m < r < 2m$). Then the output process is a correlated on-off with transition probabilities $a_{01} = p_s$ and $a_{10} = \Delta p_s/m$, where $\Delta = r - m$.

Proof: Assume that the HOL packet k departs the system at frame t. If the queue is non-empty, the next HOL packet $k+1$ is served immediately, and the interdeparture time T is exactly the service time of packet $k+1$, denoted by $S \sim \mathcal{G}_{p_s}$; otherwise, the interdeparture time is S plus the system idle time, which is one frame for $r < 2m$, *i.e.*,

$$T = \begin{cases} 1+S & \text{with probability } P_I \\ S & \text{with probability } P_B, \end{cases} \tag{33}$$

where P_I and $P_B = 1 - P_I$ are the conditional node idle and busy probabilities given a packet departure event. The pmf $\{t_j | j \geq 1\}$ of the interdeparture time T is

$$t_j = \begin{cases} P_B p_s & j = 1 \\ p_s(1-p_s)^{j-2}(1-P_B p_s) & j > 1, \end{cases} \tag{34}$$

corresponding to a correlated on-off process [24, (1),(2)]. Based on the stability condition, the average interdeparture time $\overline{T}$ is equal to the interarrival time r/m, *i.e.*,

$$\overline{T} = 1 + \frac{1}{p_s} - P_B = \frac{r}{m} \quad \Longrightarrow \quad P_B p_s = 1 + p_s - \frac{1}{\rho} = 1 - \frac{\Delta p_s}{m}. \tag{35}$$

So the transition probabilities are $a_{10} = 1 - P_B p_s = \Delta p_s / m$ and $a_{01} = p_s$. ■

Therefore, the first relay node n_1 is GI/Geom/1 system with on-off input. The queue length distribution is derived in [25], and the packet delay is shown to be geometric in [26] without calculating the parameter of the geometric distribution, which is derived in Lemma 3.6.

Lemma 3.6: Consider a discrete-time GI/Geom/1 queueing system with service rate p_s and on-off arrival, whose transition probabilities are a_{01} and a_{10}. Then, the delay is geometrically distributed with parameter

$$\xi = \frac{1-p_s}{p_s a_{10} + (1-p_s)(1-a_{01})}. \tag{36}$$

Proof: Using the delay model [24], denote the system state by the delay of the HOL packet. Negative states indicates an idle server. All probabilities of going beyond a delay -1 are included in the state -1. The transition probability Z_{jk} is

$$Z_{jk} = \begin{cases} p_s b_l(j) & j \geq 0,\ k \leq j \\ 1-p_s & j \geq 0,\ k = j+1 \text{ or } j = -1,\ k = j \\ a_{01} & j = -1,\ k = 0, \end{cases} \tag{37}$$

where $l = j - k + 1$, $b_l = t_l$ and $b_l(j) = b_l$ for $j > -1$, and $b_l(j) = \sum_{h=l}^{\infty} b_h = a_{10}(1 - a_{01})^{l-2}$ for $j = -1$. Manipulating the balance equations, the steady-state probability is $z_j = z_0 \xi^j$ $(j \geq 0)$. Since the delay probabilities involve only non-negative states (as in (31)), the delay distribution is geometric with ξ. ■

For the relay node n_1, with $a_{10} = \Delta p_s/m$ and $a_{01} = p_s$, the delay probabilities are $\pi_{jm+1} = (1-\xi)\xi^j$ at the time slot level. Therefore, the mean and variance are

$$\mu_1 = 1 + \frac{m\xi}{1-\xi} = 1 + m\varepsilon, \quad \varepsilon \triangleq \frac{\rho}{1-\rho}\frac{1-p_s}{p_s} \tag{38}$$

$$\sigma_1^2 = \frac{m^2\xi}{(1-\xi)^2} = m^2\varepsilon(1+\varepsilon). \tag{39}$$

The output of n_1 is also bursty and correlated, but too complicated to be characterized by a simple traffic model. It is stated in [26] that the output process of tandem GI/Geom/1 queues converges to a Bernoulli (geometric) process as the number of nodes goes to infinity. However, in practical wireless networks, the route is often too short for the output to converge to a geometric process. Considering that on-off is a simple yet accurate model for bursty and correlated traffic, we use the derived on-off process with $\{a_{01}, a_{10}\}$ to characterize the inputs to all relay nodes. As a matter of fact, simulation results confirm that the delays W_i $(i \geq 1)$ of the relay nodes are close to the geometric process described in Lemma 3.6.

C. Estimation of the correlation factor

The inputs to relay nodes are correlated and bursty and result in correlations between the delays W_i's. As mentioned in Section II, the correlation is generally a function of the traffic intensity $\rho = m/(rp_s)$. However, in a TDMA tandem network, the operations of the servers are successive, *i.e.*, the neighboring nodes cannot be served simultaneously. This behavior affects the correlation. It also implies that the correlation might not be solely determined by ρ. If $\rho = 0$, there is no traffic, and $W_i \equiv 0$. So, there is no correlation and $\eta = 0$. On the other hand, if $p_s = 1$, the channel is perfect, and every packet can be successfully transmitted with only one attempt regardless of ρ, *i.e.*, $W_i \equiv 1$ $(i \geq 1)$. Then, the correlation is expected to be $\eta = 0$ as well even if $\rho \neq 0$. In this sense, the correlation depends at least on two parameters ρ and p_s although ρ itself is a function of p_s since $p_s = 1$ does not necessarily lead to $\rho = 0$. From (38) and (39), it is verified that the delay W_i $(i \geq 1)$ is a function of ε. Accordingly, we assume that the correlation coefficient η is a function of ε.

It is unclear how ε impacts the correlation coefficient η. A set of simulation results (obtained by MATLAB) are provided to establish an empirical model of η. As shown in Fig. 4(a), i) the correlation is indeed negative; ii) the magnitude of the correlation can be regarded as exponentially increasing with ε. Based on the least-square principle, the curve is fitted as follows

$$\eta(\varepsilon) = x_1 + x_2 e^{-x_3 \varepsilon^{x_4}}, \quad \varepsilon = \frac{\rho}{1-\rho} \cdot \frac{1-p_s}{p_s}, \tag{40}$$

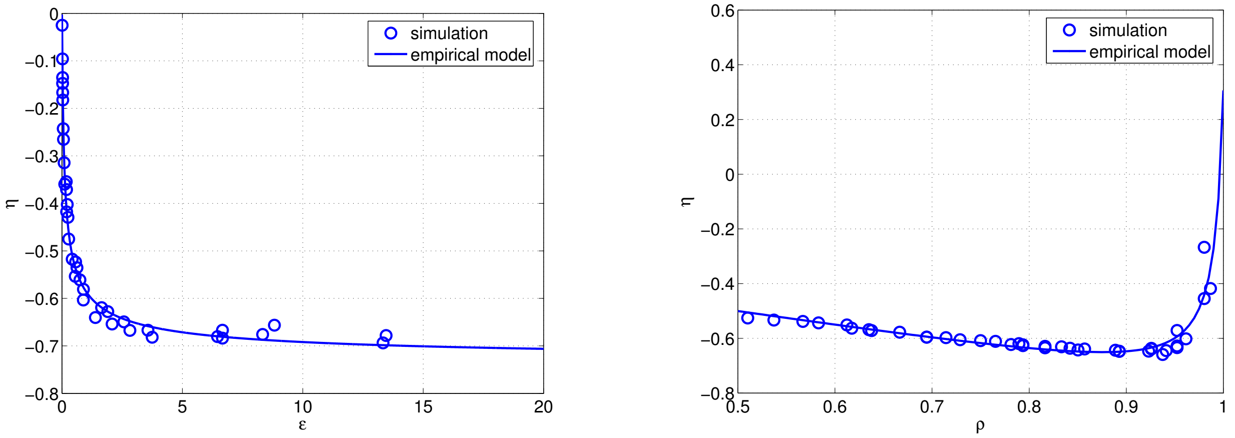
where $x_1 = -0.0023$, $x_2 = -0.7350$, $x_3 = 0.2315$, $x_4 = -0.5598$. With the empirical η and assuming that all relay nodes behave identically, the e2e delay mean and variance are further simplified from (1) to the following

$$\mu = \mu_0 + N\mu_1 \tag{41}$$

$$\sigma^2 = \sigma_0^2 + \sigma_1^2 + \eta\sigma_0\sigma_1 + (N-1)\eta\sigma_1^2 \approx \sigma_0^2 + N(1+\eta)\sigma_1^2. \tag{42}$$

Then, the tandem network can be treated as a series of *independent* discrete-time servers. The first one is the source with mean μ_0 and variance σ_0^2. The others are identical relays with mean μ_1 and scaled variance $(1+\eta)\sigma_1^2$. According to the Law of Large Numbers, the e2e delay will converge to a Gaussian distribution, and the mean and variance are linear with the number of nodes N. Simulation results in Fig. 5(a) and Fig. 3 confirm the quick convergence to a "sampled" Gaussian and the linearity even for relatively short networks.

In Fig. 3 $(m = 3, p_s = 0.8, N = 14$ for $r = 4$ and $r = 5$, respectively), we compare the mean and variances in three cases: 1) the simulated e2e delay; 2) the sum of the simulated delays of the individual nodes, *i.e.*, the e2e delay when all servers are independent; and 3) the analytical delay with mean (41) and scaled variance (42). It is obvious that, with respect to the variance,



(a) TDMA: η vs. ε (see (40))

(b) ALOHA: η vs. ρ (see (50))

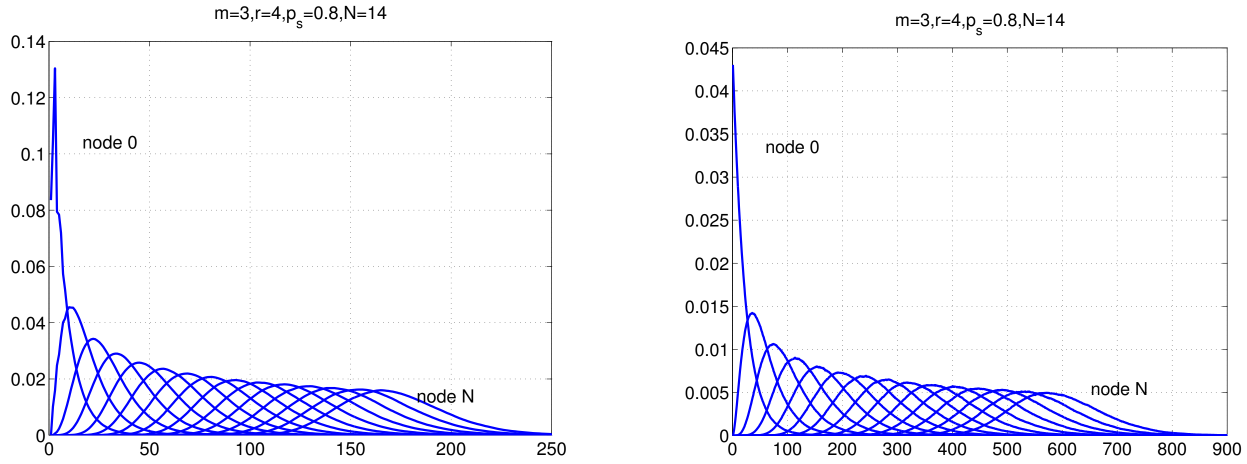
Fig. 4. The estimation of the correlation coefficient η

the correlation causes a huge gap between the real case 1) and the "independence" case 2). In other words, the "independence" assumption is not accurate for the calculation of the variance. Our analytical results (case 3)) use the empirical correlation coefficient η and are much closer to the simulated variance.

The other observation revealed in Fig. 3 is the impact of the traffic rate $1/r$. As r increases, the delay mean and variance unsurprisingly decrease. However, the delay decreases at a much faster speed than the increase of r. Comparing $r = 4$ and $r = 5$, with r increasing by 20%, the mean decreases by 70% and the variance by 87%. The reason is that the traffic intensity is determined by three parameters m, r, p_s ($\rho = m/(rp_s)$). With m and p_s fixed, a small change in r probably causes a huge change in ρ, and in μ (38) and σ^2 (39). Previous work usually assumed $\rho = 1$ thereby preventing a study of the impact of the traffic rate.

IV. Analysis of the Slotted ALOHA Network

m-phase TDMA achieves the optimal performance but also incurs a substantial amount of overhead to establish the frame structure and requires a complete cooperation between all nodes



(a) TDMA

(b) ALOHA

Fig. 5. Pmf of the packet cumulated delay with $m = 3, r = 4, p_s = 0.8, N = 14$

involved. In wireless networks, slotted ALOHA may be more practical since every node operates in a completely independent way. However, unless the traffic is light, its random and independent transmission pattern generally results in poor performance. This section analyzes the delay of ALOHA networks using a similar technique as in TDMA since the individual node is also modeled as a GI/Geom/1 system. The difference is that the ALOHA network is analyzed at the time slot level. Note that the source node n_0 is modeled as a D/Geom/1 queueing system with an integer interarrival time r. The delay distribution is derived in the following Corollary 4.1 to Theorems 3.1 and 3.2.

Corollary 4.1: Consider a D/Geom/1 queueing system with interarrival time $r \in \mathbb{N}$ and service rate s. Then the delay is $\mathcal{G}_{1-\alpha}$, where α is a positive real root of the polynomial

$$sy^r - y + 1 - s = 0. \tag{43}$$

The delay mean and variance are approximately expressed in an explicit form

$$\mu_0 = \frac{1}{1-\alpha} \approx \frac{(r-1)\rho}{2(1-\rho)}, \quad \rho = \frac{1}{sr} \tag{44}$$

$$\sigma_0^2 = \frac{\alpha}{(1-\alpha)^2} \approx \Big(\frac{(r-1)\rho}{2(1-\rho)}\Big)^2\Big(1-\frac{2(1-\rho)}{(r-1)\rho}\Big). \tag{45}$$

Proof: As a special case $m = 1$ of Theorem 3.1, the polynomial (10) is modified to (43) by replacing p_s with s and setting $m = 1$. There is only one root inside unit circle, denoted by α. Plugging $a = 1/\alpha$ and $\beta_0 = \alpha$ into (22) (Theorem 3.2), the pgf of the delay distribution is

$$G_{W_0}(z) = \frac{(1-\alpha)z}{1-\alpha z}, \tag{46}$$

which is corresponding to a geometric distribution with parameter $1-\alpha$. Using the same technique as in Proposition 3.3 but replacing Δ by $r-1$, if $r \gtrapprox 4$ and $\rho \gtrapprox 0.5$, α is well approximated by

$$\alpha \approx 1 - \frac{2(1-\rho)}{(r-1)\rho}. \tag{47}$$

Then, the delay mean μ_0 and variance σ_0^2 are approximated as in (44) and (45). ■

The output process of n_0 is more complex in ALOHA than in TDMA since the system idle time ranges from 0 to $r-1$ time slots. But it can be approximated as an on-off process, as stated in Lemma 4.2 by using a similar technique in Theorem 3.5. The proof is skipped due to the space constraints.

Lemma 4.2: Consider a D/Geom/1 queueing system with interarrival time $r \in \mathbb{N}$ and service rate s. Then, the output process is approximately an on-off with transition probabilities $a_{01} = (1-s)/((r-1)\alpha)$ and $a_{10} = (1-s)/\alpha$.

Then the relay nodes are modeled as GI/Geom/1 with on-off input $\{a_{01}, a_{10}\}$. According to Lemma 3.6, the delay W_i $(i \geq 1)$ is geometric with ξ

$$\xi = \frac{1-s}{sa_{10} + (1-s)a_{00}} = \frac{(r-1)\alpha\rho}{1-(1-(r-1)\alpha)\rho} \approx 1 - \frac{1-\rho}{r\rho - 1}. \tag{48}$$

The mean and variance are

$$\begin{aligned}\mu_1 &= \frac{1}{1-\xi} = 1 + \frac{(r-1)\rho\alpha}{1-\rho} \approx \frac{r\rho - 1}{1-\rho}.\\ \sigma_1^2 &= \frac{\xi}{(1-\xi)^2} = \frac{(r-1)\rho\alpha}{1-\rho}\big(1 + \frac{(r-1)\rho\alpha}{1-\rho}\big) \approx \Big(\frac{r\rho-1}{1-\rho}\Big)^2 - \frac{r\rho-1}{1-\rho}. \qquad (49)\end{aligned}$$

The correlations in ALOHA are different from those in TDMA because i) the nodes are not served successively; ii) the service rate is $s = p_s p_m$, which cannot reach to 1 for $p_m < 1$ even if the channel is perfect. In other words, it might not be necessary to include p_s explicitly in the calculation of the correlation. Therefore, as stated in Section II, we assume that the correlation in ALOHA is a function of the traffic intensity ρ. Through simulation results (Fig. 4(b)), η is found to be composed of two parts and characterized as follows

$$\eta = x_1 + x_2\rho + \frac{x_3}{x_4 - \rho}, \qquad (50)$$

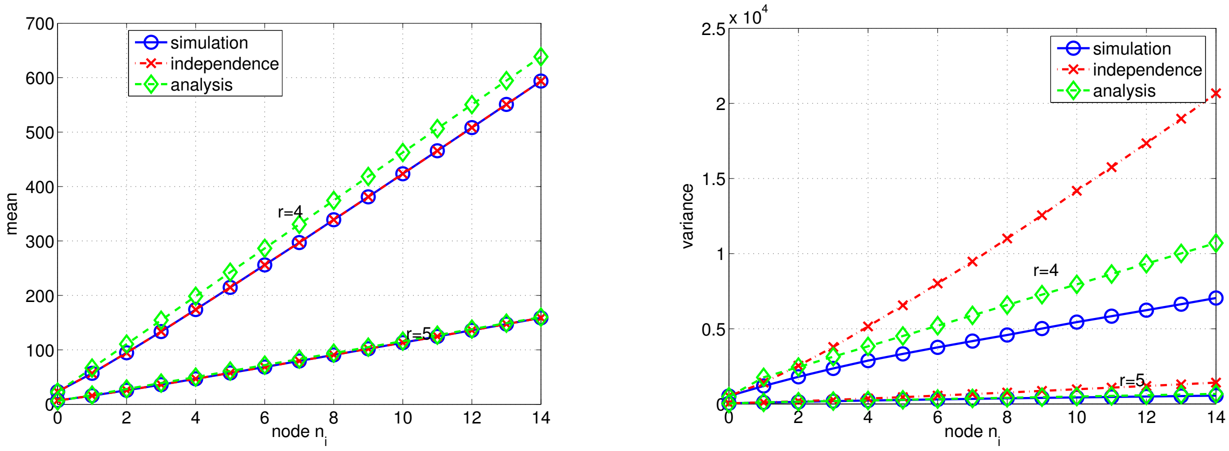
where $x_1 = -0.2483$, $x_2 = -0.5415$, $x_3 = 0.0096$, $x_4 = 1.0088$. The intuition behind this shape of $\eta(\rho)$ is: as the traffic intensity increases, the queueing delay will increase and so does the resulting correlation magnitude $|\eta|$. However, if ρ increases to 1, all delays increase to infinity. Then, $|\eta|$ might decrease rather than continue to increase. So, interestingly, a transition point $\rho \approx 0.93$ exists. Therefore, it is not suggested in ALOHA to set the traffic intensity ρ higher than this transition point because of both the increasing delay and the almost-zero correlation.

The remaining part of the ALOHA analysis is similar to TDMA. The simulated pmf (Fig. 5(b)) of the delay verifies the convergence to Gaussian. The delay mean and variance are presented in Fig. 6. The 20% increase of r causes 73% decrease of the mean and 92% of the variance.

V. Comparison

Table V lists the delay statistics of the individual node for TDMA and ALOHA. As the number of nodes N increases, the e2e delay statistics are mainly determined by those of the relay nodes. Therefore, the delay of TDMA and ALOHA can be compared as follows:

$$\frac{\mu_{\text{TDMA}}}{\mu_{\text{ALOHA}}} \approx \frac{m}{r-1} \cdot \frac{1-p_s}{p_s\alpha} \approx \frac{m}{r-1} \cdot \frac{1-p_s}{p_s} < 1, \qquad (51)$$



(a) Delay mean

(b) Delay variance

Fig. 6. Comparison of the delay performance in the ALOHA network

$$\frac{\sigma^2_{\text{TDMA}}}{\sigma^2_{\text{ALOHA}}} \approx \left(\frac{1+\eta_{\text{TDMA}}}{1+\eta_{\text{ALOHA}}}\right)\left(\frac{m}{r-1}\right)^2 \cdot \frac{1-p_s}{p_s} \cdot \left(\frac{1-p_s}{p_s} + \frac{1-\rho}{\rho}\right) < 1, \tag{52}$$

where η_{TDMA} and η_{ALOHA} are given in (40) and (50), respectively. The mean for ALOHA is $(r-1)p_s/m(1-p_s)$ times than TDMA. The ratio of the variance depends on the correlation coefficients η. The ratio of $1+\eta_{\text{TDMA}}$ to $1+\eta_{\text{ALOHA}}$ lies between 0.5 and 2.5. So, $\sigma^2_{\text{TDMA}}/\sigma^2_{\text{ALOHA}} \ll 1$ for $m \geq 3, m < r < 2m, p_s > 0.5$. As an example, for $r = 4, p_s = 0.8$ and $m = 3$ (corresponding to $p_m = 1/3$ for ALOHA), (51) and (52) shows that the mean and variance of ALOHA are 4 and 11.9 times than those of TDMA, respectively. The simulation results in Fig. 3 and Fig. 6 confirm that the delay mean of ALOHA is about four times greater than that of TDMA, while the variance is about 11 times larger. Theoretically, we proved that TDMA achieves better performance than ALOHA not only in the throughput but also in the e2e delay. This comparison is based on the assumption of equal p_s for TDMA and ALOHA. If the network were interference-limited, TDMA would have a larger p_s than ALOHA (if the other parameters are the same), which would further increase the performance gap.

In addition, with the Gaussian approximation of the e2e delay, we are able to calculate the

	Source node		Relay node	
	mean μ_0	variance σ_0^2	mean μ_1	variance σ_1^2
TDMA	$(r-m)\cdot\dfrac{\rho}{2(1-\rho)}$	$\mu_0(\mu_0-2)$	$1+m\cdot\dfrac{\rho}{1-\rho}\cdot\dfrac{1-p_s}{p_s}$	$(\mu_1-1)(\mu_1-1+m)$
ALOHA	$(r-1)\cdot\dfrac{\rho}{2(1-\rho)}$	$\mu_0(\mu_0-1)$	$1+(r-1)\cdot\dfrac{\rho}{1-\rho}\cdot\alpha$	$\mu_1(\mu_1-1)$

TABLE I

Comparison of local delays for TDMA and ALOHA

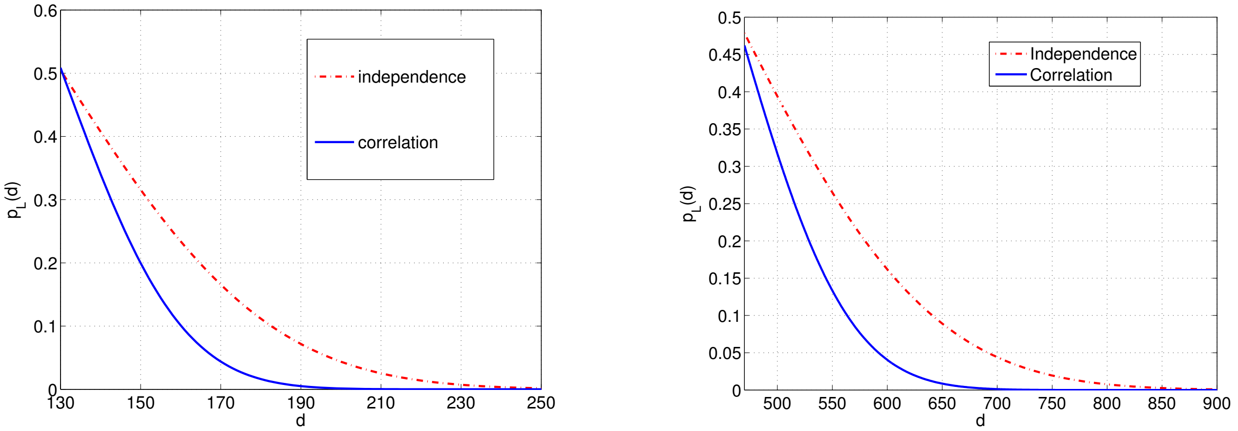
delay outage probability $p_L(d) = \Pr\{W > d\}$ for delay-sensitive applications, where d is the regarded as a hard delay bound. Given the mean μ and variance σ^2, $p_L(d)$ is

$$p_L(d) = \Pr\{W > d\} = \frac{1}{2}\Big(1 - \mathrm{erf}\big(\frac{d-\mu}{\sqrt{2}\sigma}\big)\Big). \tag{53}$$

Fig. 7 shows $p_L(d)$ for $m = 3, r = 4, p_s = 0.8, N = 10$. Compared to previous work that assumed independent W_i's, our analysis provides more accurate insight on $p_L(d)$. For example, given $d = 185$ for TDMA, we calculate $p_L(d) = 0.0093$ while the independence assumption leads to $p_L(d) = 0.09$, which is almost 10 times higher. Similarly, for ALOHA, given $d = 600$, our analysis yields $p_L(d) = 0.0087$ while the independence assumption leads to $p_L(d) = 0.0891$. Since $p_L(d)$ is an important measure for the design of real-time networks, a more accurate analysis is desired. Besides, Fig. 7 also shows how much worse ALOHA behaves in terms of the outage probability. In order to achieve $p_L(d) \leq 10\%$, ALOHA requires the delay bound $d = 570$, about 3.5 times than the delay bound $d = 160$ of TDMA.

VI. Conclusions

This paper integrates the queueing analysis of individual nodes with the MAC schemes, using queueing theory to explain the delay difference between TDMA and ALOHA. For TDMA, the resulting system is more complicated than conventional D/Geom/1 even for CBR traffic. We establish a 2-D Markov chain to derive the distributions of the queue length, delay and the



(a) TDMA

(b) ALOHA

Fig. 7. The delay outage probability $p_L(d)$ for $m = 3, r = 4, p_s = 0.8, N = 10$

output process. Unlike previous work on MAC schemes, we do not assume overly heavy traffic (*i.e.*, $\rho = 1$, such that the nodes always have packets to transmit) so that the impact of the traffic load can be explored. It is shown that a small decrease of the traffic rate may substantially improve the e2e delay.

From the perspective of queueing theory, the access delays are incorporated into the prolonged service time in TDMA and the decreased service rate in ALOHA. Given the same traffic load and original service rate, we proved that prolonging the service time by m times causes less delay than decreasing the service rate by $p_m = 1/m$ times in terms of both the mean and variance.

Regarding the e2e delay, we take into account the negative correlations between the delays since they have significant influence on the e2e delay variance and have been ignored in previous work based on the "independence" assumption. Using simulation results to establish an empirical model on the correlation coefficient, the tandem network is modeled as a series of independent servers whose variances are scaled by the correlation coefficient η. Then, we derive a more accurate expression for the e2e delay variance which has been paid little attention in the literature. The complete delay distribution leads to a delay outage probability $p_L(d)$, a critical measurement for delay-sensitive applications. Naturally, the more accurate variance results in improved tightness of $p_L(d)$.

REFERENCES

[1] H. Zhai, J. Wang, and Y. Fang, "Distributed Packet Scheduling for Multihop Flows in Ad Hoc Networks," in *IEEE Wireless Communications and Networking Conference (WCNC)*, vol. 2, Mar. 2004, pp. 1081–1086.

[2] R. Nelson and L. Kleinrock, "Spatial TDMA: A Collision-Free Multihop Channel Access Protocol," *IEEE Transactions on Communications*, vol. 33, no. 9, pp. 934–944, Sept. 1985.

[3] J. C. Arnbak and W. V. Blitterswijk, "Capacity of Slotted ALOHA in Rayleigh-Fading Channels," *IEEE Journal on Selected Areas in Communications*, vol. 5, no. 2, pp. 261–269, Feb. 1987.

[4] M. C. H. Peh, S. V. Hanly, and P. Whiting, "Random Access with Multipacket Reception over Fading Channels," in *Australian Communications Theory Workshop 2003*, Feb. 2003.

[5] F. Borgonovo and M. Zorzi, "Slotted ALOHA and CDPA: A Comparison of Channel Access Performance in Cellular Systems," *ACM Wireless Networks*, vol. 3, no. 1, pp. 43–51, Mar. 1997.

[6] Y. Yang and T.-S. P. Yum, "Delay Distributions of Slotted ALOHA and CSMA," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1846–1857, Nov. 2003.

[7] M. Zorzi and S. Pupolin, "Slotted ALOHA for High-Capacity Voice Cellular Communications," *IEEE Transactions on Vehicular Technology*, vol. 43, no. 4, pp. 1011–1021, Nov. 1994.

[8] J. A. Silvester and L. Kleinrock, "On the Capacity of Multihop Slotted ALOHA Networks with Regular Structures," *IEEE Transactions on Communications*, vol. 31, no. 8, pp. 974–982, Aug. 1983.

[9] P. Gupta and P. R. Kumar, "The Capacity of Wireless Networks," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 388–404, Mar. 2000.

[10] M. Sidi, "Tandem Packet-Radio Queueing Systems," *IEEE Transactions on Communications*, vol. 35, no. 2, pp. 246–248, Feb. 1987.

[11] C. Santivanez and I. Stavrakakis, "Study of Various TDMA Schemes for Wireless Networks in the Presence of Deadlines and Overhead," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 7, pp. 1284–1304, July 1999.

[12] L. G. Roberts, "ALOHA Packet System With and Without Slots and Capture," *ACM Sigcomm Computer Communication Review*, vol. 5, no. 2, pp. 28–42, Apr. 1975.

[13] J. A. Morrison, "Two Discrete-Time Queues in Tandem," *IEEE Transactions on Communications*, vol. 27, no. 3, pp. 563–573, Mar. 1979.

[14] J. Hsu and P. J. Burke, "Behavior of Tandem Buffers with Geometric Input and Markovian Output," *IEEE Transactions on Communications*, vol. 24, no. 3, pp. 358 – 361, Mar. 1976.

[15] M. J. Neely, "Exact Queueing Analysis of Discrete Time Tandems with Arbitrary Arrival Processes," in *IEEE International Conference on Communications (ICC'03)*, vol. 4, 2003, pp. 2221 – 2225.

[16] N. Gulpinar, P. Harrison, and B. Rustem, "Mean-Variance Optimization of Response Time in a Tandem M/GI/1 Router Network with Batch Arrivals," in *The* 3*rd International Working Conference on Performance Modeling and Evaluation of Heterogeneous Networks*, July 2005.

[17] P. Jacquet, A. M. Naimi, and G. Rodolakis, "Routing on Asymptotic Delays in IEEE 802.11 Wireless Ad Hoc Networks," in 1*st Workshop on Resource Allocation in Wireless NETworks (RAWNET) 2005*, Apr. 2005.

[18] F. Eshghi, A. K. Elhakeem, and Y. R. Shayan, "Performance Evaluation of Multihop Ad Hoc WLANs," *IEEE Communications Magazine*, pp. 107–115, Mar. 2005.

[19] I. Mitrani and R. Chakka, "Spectral Expansion Solution for A Class of Markov Models: Application and Comparison with the Matrix-geometric Method," *Performance Evaluation*, vol. 23, no. 3, pp. 241–260, Sept. 1995.

[20] A. Cohn, "Ueber die Anzahl der Wurzeln einer algebraischen Gleichung in einem Kreise," pp. 112–124, 1921, Available at http://dz-srv1.sub.uni-goettingen.de/sub/digbib/pdftermsconditions?did=%D42634&p=6 .

[21] J. J. Hunter, *Mathematical Techniques of Applied Probability*. Academic Press, Oct. 1983, ISBN:0123618029.

[22] M. L. Chaudhry, U. C. Gupta, and J. G. C. Templeton, "On the Relations Among the Distributions at Different Epochs for Discrete-Time GI/Geom/1 Queues," *Operations Research Letters*, vol. 18, pp. 247–255, 1996.

[23] Http://functions.wolfram.com/ElementaryFunctions/Log/29/.

[24] K. K. Lee and S. T. Chanson, "Packet Loss Probability for Bursty Wireless Real-Time Traffic Through Delay Model," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 3, pp. 929–938, May 2004.

[25] I. Elhanany, M. Kahane, and D. Sadot, "On Uniformly Distributed On/Off Arrivals in Virtual Output Queued Switches with Geometric Service Times," in *IEEE International Conference on Communications (ICC'03)*, vol. 1, May 2003, pp. 173–177.

[26] B. Prabhakar and R. Gallager, "Entropy and the Timing Capacity of Discrete Queues," *IEEE Transactions on Information Theory*, vol. 49, no. 2, pp. 357–370, Feb. 2003.