MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Queueing Delay Analysis of Mixed Traffic in Time Sensitive Networks

Jurdi, Rebal; Guo, Jianlin; Kim, Kyeong Jin; Orlik, Philip V.; Nagai, Yukimasa

TR2021-122 October 20, 2021

Abstract

In many emerging Time Sensitive Networking (TSN) applications such as industrial control and automotive, periodic traffic with end-to-end latency constraints is buffered with non-mission-critical aperiodic traffic. As queueing delay is a central component of end-to-end latency, we study the effect of aperiodic traffic on the queueing delay of periodic traffic via characterizing the probability distribution of queue size and delay in an (M+D)/M/1 queue, a queue with Poisson and periodic inputs, infinite waiting capacity and a single memoryless server. Since obtaining the closed form distributions of the queue size and delay is intractable, we approximate the behavior of the (M+D)/M/1 queue when the service and arrival rates are close. We use a Markov chain with a quasitoeplitz matrix, enabling us to use an existing technique to study this class of Markov chains. We determine the transition matrix of the Markov chain, investigate the setting in which the approximation works well, and compute the stationary distribution. We plot and compare the stationary distributions of queue size and delay between the (M+D)/M/1 queue and our model; we observe that our model is a favorable match.

International Conference on Intelligent Manufacturing and Automation Engineering (ICIMA)

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Queueing Delay Analysis of Mixed Traffic in Time Sensitive Networks

Rebal Jurdi*‡, Jianlin Guo*, Kyeong Jin Kim*, Philip Orlik* and Yukimasa Nagai†
*Mitsubishi Electric Research Laboratories
†Mitsubishi Electric Corporation IT R&D Center
‡The University of Texas at Austin

***Abstract*—In many emerging Time Sensitive Networking (TSN) applications such as industrial control and automotive, periodic traffic with end-to-end latency constraints is buffered with non-mission-critical aperiodic traffic. As queueing delay is a central component of end-to-end latency, we study the effect of aperiodic traffic on the queueing delay of periodic traffic via characterizing the probability distribution of queue size and delay in an $(M+D)/M/1$ queue, a queue with Poisson and periodic inputs, infinite waiting capacity and a single memoryless server. Since obtaining the closed form distributions of the queue size and delay is intractable, we approximate the behavior of the $(M+D)/M/1$ queue when the service and arrival rates are close. We use a Markov chain with a quasitoeplitz matrix, enabling us to use an existing technique to study this class of Markov chains. We determine the transition matrix of the Markov chain, investigate the setting in which the approximation works well, and compute the stationary distribution. We plot and compare the stationary distributions of queue size and delay between the $(M+D)/M/1$ queue and our model; we observe that our model is a favorable match.**

***Index Terms*—Time sensitive networking, queueing delay, mixed traffic, (M+D)/M/1 queue, QoS.**

I. Introduction

Time sensitive networking (TSN) is an emerging set of technologies targeting real time applications such as industrial control, industrial automation and automotive [1]. IEEE and IEC are collaboratively developing TSN standards and profiles. TSN builds timing, synchronization, and scheduling functions on top of the IP stack to enable deterministic communication across the network [2].

There are two classes of traffic in TSN networks in regard to communication latency, TSN data and non-TSN data. TSN data is time-triggered data with stringent constraints on communication latency. Consider for example a closed-loop (feedback) industrial control system made of controllers, actuators, physical plants, and sensors. The sensors continuously measure the state of an underlying process or plant by collecting state variables, such as temperature and pressure, and report these measurements to the controllers. The controllers then compute the desired state and dispatch commands to the actuators to respond accordingly. For real-time applications, this collect-compute-command cycle must complete within a millisecond [1], further constraining the end-to-end delay from sensor to controller and from controller to actuator. Non-TSN data, however, is non-time-critical traffic carried by best-effort delivery services and is thus aperiodic. In the context of the industrial control example, the controllers communicate with devices not involved in the control loop, some of which are physically outside the industrial premises. Since packets from the two sources of traffic traverse the same network stack, aperiodic traffic is buffered with periodic traffic, resulting in increased queuing delay of periodic TSN traffic, and thus increased end-to-end delay.

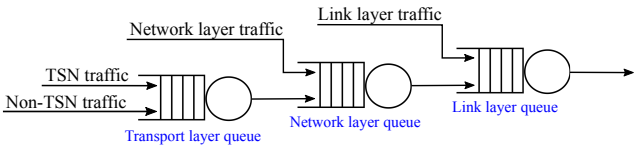


Fig. 1. The multi-level queue abstraction of the network stack of a TSN device. Periodic TSN packets generated at the application layer are buffered with aperiodic non-TSN packets in a transport layer queue. TSN packets have to traverse subsequent queues at lower layers until they are finally transmitted.

Another TSN application where periodic traffic is mixed with aperiodic traffic is synchronization where a device with a reference clock (master) synchronizes other devices (slaves) by periodically exchanging an application layer handshake [3]. The slaves correct their clocks by estimating their offset from the master clock and subtracting the propagation delay from that offset, but other delays remain unaccounted for, namely the queuing delay at the different layers of the IP stack. With aperiodic data messages sharing the same buffer with synchronization messages, queuing delay is only expected to increase.

From the application perspective, the total delay through the network stack is the sum of delays across the individual buffers at the different layers. While these buffers are typically modeled by a sequence of queues (see Fig. 1). To the best of our knowledge, there are no existing models for queues with both periodic and aperiodic traffic. Incidentally, studying queuing delays in TSN applications has led us to a more fundamental queuing-theoretic problem: characterizing the probability distribution of delay in a queue that buffers periodic packets with aperiodic packets. The simplest of such a queue would have a single server and infinite capacity, i.e. no-blockages. Periodic traffic can be modeled by a renewal process (interarrivals are IID) with deterministic interarrivals, and aperiodic traffic by a Poisson arrival process. Even though aperiodic traffic has been shown to exhibit a heavy-tail distribution, we choose the

The work of Rebal Jurdi was done during his internship at Mitsubishi Electric Research Laboratories (MERL).

Poisson model for tractability; we also choose a memoryless server for the same reason. This queue would be referred to as $(M+D)/M/1$ in Kendall's notation. To the best of our knowledge, this queue has not been previously analyzed. In the coming sections, we frame this paper within work on end-to-end delay analysis and within queuing-theoretic frameworks in general.

The paper is organized as follows. Sec. II frames this work within related prior work. Sec. III introduces the queuing model and derives a Markov-chain approximation. Sec. V analyzes the stationary distribution of the queue size and characterizes the delay. Sec. VI presents numerical results to verify the model and analysis. Sec. VII concludes the paper.

II. Related Work

This paper propose a queuing theoretic model to obtain the probability distribution of queuing delay of real-time data, a key component of end-to-end delay in TSN applications. Delay models in data networks have been heavily researched [4]. Early work used a network of queues in tandem to model a sequence of buffers at different nodes in the network or at different layers of the network stack of a single node [5]. The tandem queue model has thrived and continues to appear in recent work, for e.g. work investigating queues with packet delivery deadlines [6], and work investigating the end-to-end delay across multiple hops in a wireless network [7]. Using the tandem queue model to represent a sequences of nodes abstracts the network stack within every node as a single queue. Our paper follows this convention and describe the network stack by a single queue. The tandem queuing model, assumes Markov arrivals (e.g. Poisson) and Markov servers, enabling the description of queue sizes as Markov chains and making for a tractable analysis. We, however, assume a superposition of Poisson and periodic arrivals.

When the size of a queue is a continuous-time Markov chain, its distribution π can be analyzed tractably. If the size is a non-Markov process, it is customary to look at an embedded discrete-time Markov chain whose state transitions are aligned with those of the original, continuous-time process [8]. The size distribution is then determined in 2 steps: finding the transition matrix P of the embedded chain, and solving the *global balance equation* $\pi = \pi P$ [8]. While the first step is straightforward, the second is generally challenging especially when the state space of the Markov chain is infinite (finding an eigenvector of an infinite matrix). There often are simplifying methods in some special cases, e.g. the finite difference method, the Wiener-Hopf method, and the Riemann-Hilbert factorization. If these methods fail to apply, the fallback method is to recursively solve the global balance equations. More complicated queuing models consider different classes of packets grouped according to priority, destination, service requirements, or routes through the network. Every class has its own arrival process, and these processes are mostly taken to be Poisson and independent of one another. We consider two classes according to their arrival processes: aperiodic data represented by a Poisson process, and periodic data represented by a renewal process with deterministic interarrivals. To the best of our knowledge, the closest existing models are the $G/M/1$ queue where the arrivals form a renewal process [4], and another queue admitting a superposition of IID periodic streams of data [9]. The size of the $(M+D)/M/1$ is not a Markov chain. Additionally, its embedded Markov chain yields a set of global balance equations that are intractable. Therefore, we approximate the embedded Markov chain with another chain that has a *quasitoeplitz* transition matrix. Markov chains with quasitoeplitz transition matrices have been investigated in [10], where the steady-state distribution is determined through its probability generating function (PGF) by solving a Riemann boundary value problem on the unit circle. Extensions to transition matrices with block-repeating entries can be found in [11] and [12]. We use the method described in [10] to find the stationary distribution. We note that this approximation is valid when the service rate is close to (but greater than) the arrival rate. We verify the validity of this approximation by plotting both the stationary distributions of queue size and delay for the $(M+D)/M/1$ queue against those of our model under different parameter combinations; the results show that our model favorably matches the original queue.

III. Queue Model

Consider an infinite-capacity, single-server queue with an exponential service distribution of rate μ. Also consider an arrival process $A(t)$ that is the superposition of a homogeneous Poisson point process $A_S(t)$ with density λ and a periodic point process $A_T(t)$ with period T, i.e.

$$A(t) = A_S(t) + A_T(t). \tag{1}$$

We adopt Kendall's notation and refer to this queue as $(M+D)/M/1$. The processes $A_S(t)$ and $A_T(t)$ are renewal processes, i.e. the interarrival times of each process are positive and IID random variables; interarrival times of $A_S(t)$ are exponential of rate λ and interarrival times of $A_T(t)$ are deterministic and equal to T. The stationary distribution of queue size for queues with renewal inputs has been analyzed, closed-form expressions of the distribution have been determined for both a Poisson input a and periodic input, and a near-closed-form expression for a renewal input with arbitrary (general) distribution has been outlined [13]. The input process $A(t)$ that we consider, however, is not a renewal process [14], so these results do not apply. Additionally, to the best of our knowledge, there is no prior work on the $(M+D)/M/1$ queue, and no general recipes to analyze the behavior of a queue with Poisson and generally-distributed arrivals.

Now, let $Q(t)$ be the size of the queue (the number of packets) at an arbitrary time t. Consider the discrete-time process Q_n representing the size of the queue right before the t_n, i.e. $Q_n = Q(t_n - 0)$. The process Q_n is a Markov chain as it can be described by the following recurrence relation:

$$Q_{n+1} = f(Q_n + 1), \tag{2}$$

where $f(i)$ is the size of an $M/M/1$ queue at time T when it initially has a size of i. The transition from an initial size of i to a size of $j = f(i+1)$ after T time units has a probability

$$\begin{aligned} q_{ij} = {} & e^{-(\lambda+\mu)T} \rho^{\frac{j-i-1}{2}} I_{j-i-1}(2T\sqrt{\lambda\mu}) \\ & + e^{-(\lambda+\mu)T} \rho^{\frac{j-i-2}{2}} I_{j+i+2}(2T\sqrt{\lambda\mu}) \\ & + e^{-(\lambda+\mu)T}(1-\rho)\rho^k \sum_{k=j+i+3}^{\infty} \rho^{-k/2} I_j(2T\sqrt{\lambda\mu}). \end{aligned} \tag{3}$$

By defining

$$u_{ij} = e^{-(\lambda+\mu)T} \rho^{\frac{j-i-1}{2}} I_{j-i-1}(2T\sqrt{\lambda\mu}) \tag{4}$$

and

$$\begin{aligned} v_{ij} = {} & e^{-(\lambda+\mu)T} \rho^{\frac{j-i-2}{2}} I_{j+i+2}(2T\sqrt{\lambda\mu}) \\ & + e^{-(\lambda+\mu)T}(1-\rho)\rho^k \sum_{k=j+i+3}^{\infty} \rho^{-k/2} I_j(2T\sqrt{\lambda\mu}), \end{aligned} \tag{5}$$

the transition matrix of Q_n can accordingly be expressed as the sum of an infinite Toeplitz matrix $U = [u_{ij}]$, a matrix with constant diagonals, and an infinite Hankel matrix $V = [v_{ij}]$, a matrix with constant skew diagonals. With the current structure of the transition matrix, the problem of determining the stationary distribution appears to be intractable. Eliminating the Hankel matrix can serve our purpose, as the stationary distribution for a Markov chain with an infinite Toeplitiz transition matrix can in many cases be determined through its probability generating function (PGF).

IV. Approximation

We now turn to approximating the Markov chain Q_n with a Markov chain with a more structured transition matrix. This allows us to characterize the stationary distribution and thus the distribution of queueing delays.

Let t_n be the epoch of the nth periodic arrival, i.e. $t_n = nT$. Let A_n and B_n be random variables representing the number of arrivals and *potential* departures in the interval (t_n, t_{n+1}). We use the term "potential", in line with convention [13] and [8], to indicate that there might not be B_n actual departures if B_n is greater than the number of customer just after t_n. Since the length of (t_n, t_{n+1}) is T, A_n and B_n are Poisson random variables with rate parameters λT and μT. We aim to approximate Q_n by the discrete-time process $X_n = X(t_n - 0)$ given by the recurrence relation

$$X_{n+1} = \begin{cases} X_n + 1 + A_n - B_n & \text{if } X_n + 1 + A_n - B_n > 0, \\ 0 & \text{if } X_n + 1 + A_n - B_n \le 0. \end{cases} \tag{6}$$

The process X_n is a Markov chain, as the next state is a function of the current state and other random variables that are independent of current and past states. There is one caveat to (6): The order of arrivals and departures matters; if all departures follow all arrivals, or if arrivals and departures are interleaved, then (6) holds. This approximation is reasonable under two considerations: (1) The Poisson arrival rate λ is close to but less than the service rate μ, and (2) the period of periodic arrivals T is small enough so that there could practically be at most one arrival and at most one departure. In other words, there is the following tradeoff: T should be high enough so that $\lambda \approx \mu$, but T should be small enough so that $A_n \le 1$ and $B_n \le 1$ with a high probability. We give a short mathematical discussion. Suppose we are given λ and T and want to choose the best μ so that $\mathbb{P}[A_n \le 1 \text{ and } B_n \le 1] \approx 1$. The stability of the queue, however, requires that $\mu T > \lambda T + 1$. This allows us to define the difference $x = \mu T - \lambda T - 1 > 0$ and the function

$$\phi_\lambda(x) = \mathbb{P}[A_n \le 1 \text{ and } B_n \le 1] \tag{7}$$

$$= (x + 2 + \lambda T)(1 + \lambda T) e^{-2\lambda T - 1 - x} \tag{8}$$

which has its maximum at $x = 0$ by noting that $e^x = 1 + x + o(1)$. This implies that the Markov chain X_n best approximates Q_n when $\mu T \gtrapprox \lambda T + 1$. We verify our claim in Sec. VI.

Poisson interarrivals and exponential service times allow us to write the recurrence relationship of (6) that gives the evolution of the queue state in terms of the difference of two Poisson random variables, A_n for arrivals and B_n for departures. This is possible only because of the memoryless property of the exponential distribution which allows us to express the number of arrivals and services in a given observation window as a function of the *length* of that window. To extend our model to general (arbitrary) interarrival and service distributions, a different embedded Markov chain ought to be considered.

V. Delay Characterization

The first step toward characterizing the delay in a queue is determining the stationary distribution of its size.

A. Stationary distribution

We proceed to determine the transition matrix $P = [p_{ij}]$, where $p_{ij} = \mathbb{P}[X_{n+1} = j \mid X_n = i]$. We have

$$p_{ij} = \begin{cases} P(A_n - B_n = j - i - 1) & \text{if } j > 0, \\ P(A_n - B_n \le -i - 1) & \text{if } j = 0. \end{cases} \tag{9}$$

The difference $A_n - B_n$ is a difference of two Poisson random variables and it has the known *Skellam* distribution, hence we can write

$$p_{ij} = \begin{cases} e^{-(\lambda+\mu)T} \left(\frac{\lambda}{\mu}\right)^{\frac{j-i-1}{2}} I_{j-i-1}(2T\sqrt{\lambda\mu}) & \text{if } j > 0, \\ e^{-(\lambda+\mu)T} \sum_{m=i+1}^{\infty} \left(\frac{\lambda}{\mu}\right)^{-\frac{m}{2}} I_m(2T\sqrt{\lambda\mu}) & \text{if } j = 0, \end{cases} \tag{10}$$

where $I_m(\cdot)$ is the modified Bessel function of the first kind and integer order m.

Now we show that the transition probability p_{ij} is a function of the difference $i - j$ when $j > 0$. Let $\rho = \lambda/\mu$ be the standard *utilization* parameter, and $h_k = p_{i,i+k}$ the probability of adding k packets to the queue, then

$$h_k = e^{-(\lambda+\mu)T} \rho^{\frac{k-1}{2}} I_{k-1}(2T\sqrt{\lambda\mu}). \tag{11}$$

and $h_{-k} = p_{i,i-k}$ the probability of removing $k > 0$ packets from the queue. Using the fact that $I_{-k}(x) = I_k(x)$ for integer

order k, the probability of removing $k>0$ packets from the queue $p_{i,i-k}$ equals

$$h_{-k} = e^{-(\lambda+\mu)T}\ \rho^{-\frac{k+1}{2}} I_{k+1}(2T\sqrt{\lambda\mu}). \tag{12}$$

The probability of transitioning to the empty state from state i, $p_{i,0}$, is equal to

$$h_i^{\infty} = \sum_{m=i+1}^{\infty} h_{-(m-1)} = \sum_{k=i}^{\infty} h_{-k} = \sum_{k=-\infty}^{-i} h_k. \tag{13}$$

Finally, the transition matrix

$$P = \begin{bmatrix} h_0^{\infty} & h_1 & h_2 & h_3 & h_4 & \dots \\ h_1^{\infty} & h_0 & h_1 & h_2 & h_3 & \dots \\ h_2^{\infty} & h_{-1} & h_0 & h_1 & h_2 & \dots \\ h_3^{\infty} & h_{-2} & h_{-1} & h_0 & h_1 & \dots \\ h_4^{\infty} & h_{-3} & h_{-2} & h_{-1} & h_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{14}$$

A Markov chain is regular if some power of its transition matrix has only positive entries. Since P has only positive entries, X_n is regular. Therefore, it has a unique steady-state distribution π determined by the fixed-point equation

$$\pi = \pi P, \tag{15}$$

i.e. π is the probability mass function (PMF) of $X \triangleq X_{\infty}$. To guarantee that π is a probability vector, the following constraint must be added:

$$\sum_{k=0}^{\infty} \pi_k = 1. \tag{16}$$

Additionally, X_n has a *quasitoeplitz* transition matrix, i.e. there is an $i^* \geq 0$ such that for all $i \geq i^*$

$$p_{ij} = \begin{cases} h_{j-i} & \text{if } j>0, \\ \sum_{k=-\infty}^{-i} h_k & \text{if } j=0. \end{cases} \tag{17}$$

Comparing the entries of P in (14) to (17), we observe that we can take $i^*=0$. To find π, we follow the procedure described in [10] which we outline next.

First, we introduce the PGFs

$$\Pi(z) = \sum_{k=0}^{\infty} \pi_k z^k, \quad z \in \overline{\Gamma}^{+}, \tag{18}$$

$$H(z) = \sum_{k=-\infty}^{\infty} h_k z^k,\ z \neq 0, \tag{19}$$

where Γ is the unit circle $\{|z|=1\}$, Γ^{+} is the interior $\{|z|<1\}$, and $\overline{\Gamma}^{+}$ is the disc $\{|z|\leq 1\}$.

We define the operator $\{\cdot\}_{+}$ by its action on a Laurent series $f(z)=\sum_{k=-\infty}^{\infty} f_k z^k$ as

$$\{f(z)\}_{+} = \sum_{k=0}^{\infty} f_k z^k.$$

When only $f(z)$ is given, then the coefficients $\{f_k\}$ can be determined by Cauchy's integral formula. Accordingly,

$$f_k = \frac{1}{2\pi i} \oint_{\gamma} \frac{f(z)}{z^{k+1}} \mathrm{d}z, \tag{20}$$

where γ is a simple counterclockwise contour encircling $z=0$ lying within the convergence region of $f(z)$. We also define

$$R(z) = \frac{1-H(z)}{1-z^{-1}} \tag{21}$$

and

$$R^{+}(z) = \exp\left(-\{\ln R(z)\}_{+}\right) \tag{22}$$

$$= \exp\left(-\sum_{k=0}^{\infty} a_k z^k\right). \tag{23}$$

Finally, [10][eq. (40)] gives $\Pi(z)$ whenever $H'(1)<0$ as

$$\Pi(z) = \frac{R^{+}(z)}{R^{+}(1)}. \tag{24}$$

We can extend the continuity of $R^{+}(z)$ at $z=1$ by defining

$$R^{+}(1) = \lim_{z\to 1} R^{+}(z) = \exp\left(-\sum_{k=0}^{\infty} a_k\right).$$

The function $H(z)$ is the PGF of a Skellam random variable shifted by 1, and it is given as

$$H(z) = \sum_{k=-\infty}^{\infty} e^{-(\lambda+\mu)T}\ \rho^{\frac{k-1}{2}} I_{k-1}(2T\sqrt{\lambda\mu})\ z^k \tag{25}$$

$$= z \sum_{m=-\infty}^{\infty} e^{-(\lambda+\mu)T}\ I_m(2T\sqrt{\lambda\mu})\ \left(z\sqrt{\rho}\right)^m \tag{26}$$

$$= z\ \exp\left(\lambda T z + \mu T z^{-1} - (\lambda+\mu)T\right), \tag{27}$$

where the last equation is due to the fact that the generating function of $I_m(t)$ is $\exp((t+t^{-1})z/2)$ for $z\neq 0$. Its derivative

$$H'(z) = \frac{\lambda T z^2 + z - \mu T}{z}\ \exp\left(\lambda T z + \mu T z^{-1} - (\lambda+\mu)T\right), \tag{28}$$

and the constraint that $H'(1)<0$ translates to the constraint

$$\mu > \lambda + \frac{1}{T} \tag{29}$$

which is exactly the stability condition of the queue. To determine $R^{+}(z)$, we need to determine $\{\ln R(z)\}_{+}$, and the first step is selecting an appropriate contour γ along which to integrate $\ln R(z)/z^{k+1}$ according to (20). We write $R(z)$ as

$$R(z) = \frac{z(1-H(z))}{z-1}. \tag{30}$$

To determine where $\ln R(z)$ is analytic, we look at two potential singularities of $R(z)$ and $\ln R(z)$, $z=0$ and $z=1$. According to (19), $H(z)$ has an essential singularity at $z=0$ because $h_k \neq 0$ for all $k<0$; thus $1-H(z)$ also has an essential singularity at $z=0$. Therefore, $\ln R(z)$ is not analytic at $z=0$. Since h_k is a PMF, $H(1)=1$, but $H'(1)<0\neq 1$. In other words, $1-H(1)=0$, but

$$\left.\frac{\mathrm{d}(1-H(z))}{\mathrm{d}z}\right|_{z=1} \neq 0, \tag{31}$$

so that $1-H(z)$ has a simple zero at $z=1$ while $R(z)$ does not. Therefore, $\ln R(z)$ is analytic at $z=1$. Since, additionally, $H(z)$ is analytic on the unit circle Γ, we can take γ to be Γ, which enables us to express the coefficients $\{a_k\}$ of the series in (23) as an inverse discrete-time Fourier transform (DTFT), namely

$$a_k = \frac{1}{2\pi}\int_0^{2\pi} \ln\left(\frac{1-H(e^{i\Omega})}{1-e^{-i\Omega}}\right) e^{ik\Omega}\,\mathrm{d}\Omega. \tag{32}$$

Using the McLaurin series expansion of the exponential function in (23), we compute

$$R^+(z) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!}\left(\sum_{k=0}^{\infty} a_k z^k\right)^n, \tag{33}$$

where the inner sum $\{\ln R(z)\}_+$ can further be further expressed as the series

$$\sum_{k=0}^{\infty} b_k^{(n)} z^k, \tag{34}$$

with $b_0^{(n)} = a_0^n$, when $a_0 \neq 0$, and the sequence $\{b_k^{(n)}\}$ can be determined recursively through the relationship

$$b_k^{(n)} = \frac{1}{ka_0}\sum_{j=1}^{k}(nj+j-k)a_j\, b_{k-j}^{(n)}. \tag{35}$$

Plugging in the expression of $R^+(z)$ in (24), we conclude that

$$\pi_k = \frac{1}{R^+(1)}\sum_{n=0}^{\infty}\frac{(-1)^n}{n!} b_k^{(n)}. \tag{36}$$

B. Delay Distribution

We can now determine the delay D of a packet which equals the waiting time plus the service time. Let $\{S_n\}$ be a set of IID exponential service times, we can express D through its cummulative distribution function

$$F_D(t) = \mathbb{P}\left[\sum_{n=1}^{X+1} S_n \leq t\right] \tag{37}$$

$$= \sum_{k=0}^{\infty} \pi_k \frac{\mu^{k+1}}{k!} t^k e^{-\mu t} \tag{38}$$

by using the law of total expectation.

VI. Numerical Results

In this section, we compare the probability of queue size and delay for six queues with identical service processes:

1) An $(M+D)/M/1$ queue described by the process Q_n as in Sec. III. This is the reference with which we compare the rest of the queues.
2) Our Markov-chain approximation X_n.
3) An $M/M/1$ queue with an arrival density λ. This queue has only a Poisson input.
4) An $M/M/1$ queue with an arrival density $\lambda + 1/T$.
5) A $D/M/1$ queue with arrivals of period T. This queue has only a periodic input.
6) A $D/M/1$ queue with arrivals of period $T/(1+\lambda T)$.

We expect that the queue size probability curves for the $M/M/1$ queues have a tight spread against probability curve of the $(M+D)/M/1$ benchmark when $T \to \infty$, and this is indeed the case. Similarly, we expect and can verify through simulation that the probability curves for the $D/M/1$ are close to benchmark curve when $\lambda \to 0$. The queue size probabilities π_k^M and π_k^D for an $M/M/1$ queue with an arrival density λ and a $D/M/1$ queue with an arrival period T have very simple closed-form expressions [8],

$$\pi_k^M = (1-\rho)\rho^k, \tag{39}$$

$$\pi_k^D = \begin{cases} 0 & \text{if } k=0, \\ (1-\delta)\delta^{k-1} & \text{if } k \geq 1. \end{cases} \tag{40}$$

where $\rho = \lambda/\mu$ as defined earlier and δ satisfies the equation

$$\delta = e^{-\mu T(1-\delta)}. \tag{41}$$

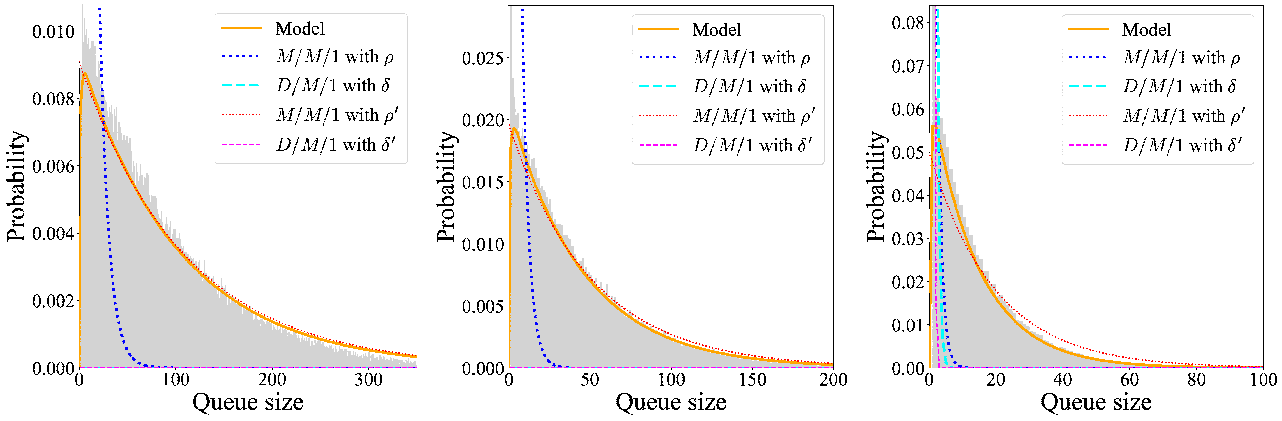
We similarly define ρ' and δ' for an arrival density of $\lambda + 1/T$ and an arrival period of $T/(1+\lambda T)$, respectively. These are only two corner cases where the simple closed-form stationary distribution expressions can be used to approximate the stationary distribution of the more complicated $(M+D)/M/1$ queue. When the two sources of traffics are of the same order, we turn to our model for an approximation. Accordingly, we consider 3 (T,λ,μ) parameter combinations with $x = \mu T - \lambda T - 1 = 0.05,\ 0.09,$ and 0.1.

Fig. 2 shows the probability of queue size for six queues. We observe that our model matches the benchmark more accurately as the difference x tends to 0, which agrees with the claim that we made earlier in Sec. III. In other words, the stationary distribution of our model favorably matches that of the $(M+D)/M/1$ queue when $\mu T \gtrapprox \lambda T + 1$. We also observe that our chosen $M/M/1$ and $D/M/1$ queues fail to match the benchmark behavior, which is the expected outcome for the chosen parameter combinations.

Fig. 3 demonstrates the probability of delay (in the unit of service rate time) for $(M+D)/M/1$ queue and our Markov-chain approximation X_n. Our Markov chain model showing a good match with the $(M+D)/M/1$ queue.

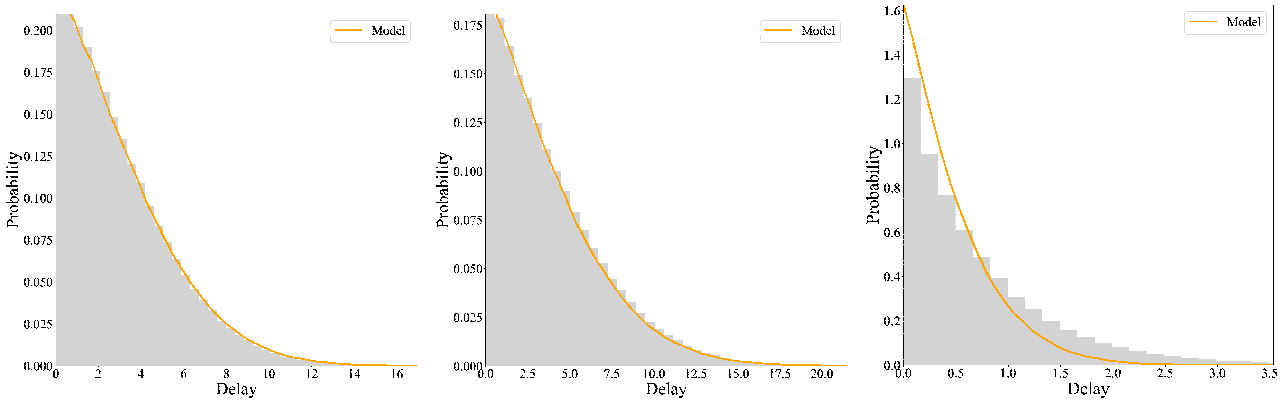
VII. Conclusion

We have proposed a Markov chain model to approximate the behavior of an $(M+D)/M/1$ queue under high utilization, i.e. when the ratio of arrival rate to service rate is high. For an analytical handle on queue size and delay, we determined the stationary distribution of queue size from which the distribution of delay could be readily determined. This approximation is valid under high utilization because the probability of having at most one Poisson arrival and at most one departure every period is high, albeit far from 1. This motivates truncating the distribution of arrivals and that of departures per period, leading to a *band* transition matrix and a potentially more tractable analysis. With a model to study a queue with periodic and aperiodic inputs, we have characterized the queueing delay of time-sensitive traffic which is a central component of end-to-end latency.



(a) $T = 1.1,\ \lambda = 9,\ \mu = 10$ (b) $T = 1,\ \lambda = 4,\ \mu = 5.1$ (c) $T = 0.1,\ \lambda = 10,\ \mu = 20$

Fig. 2. The stationary distribution of queue size for the 5 queues and our Markov chain model.



(a) $T = 1.1,\ \lambda = 9,\ \mu = 10$ (b) $T = 1,\ \lambda = 4,\ \mu = 5.1$ (c) $T = 0.1,\ \lambda = 10,\ \mu = 20$

Fig. 3. The distribution of delay (waiting time plus service time).

REFERENCES

[1] A. Nasrallah, A. Thyagaturu, Z. Alharbi, C. Wang, X. Shao, M. Reisslein, and H. ElBakoury, "Ultra-Low Latency (ULL) Networks: A Comprehensive Survey Covering the IEEE TSN Standard and Related ULL Research," *ArXiv e-prints*, Mar. 2018.

[2] "Time-sensitive networking: A technical introduction," Cisco, Tech. Rep. C11-738950, 2017. [Online]. Available: https://www.cisco.com/c/dam/en/us/solutions/collateral/industry-solutions/white-paper-c11-738950.pdf

[3] M. Lvesque and D. Tipper, "A survey of clock synchronization over packet-switched networks," *IEEE Commun. Surv. Tutor*, vol. 18, no. 4, pp. 2926–2947, 2016.

[4] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Prentice Hall, 1992.

[5] W. Chu, G. Fayolle, and D. Hibbits, "An analysis of a tandem queueing system for flow control in computer networks," *IEEE Trans. on Comput.*, vol. C-30, no. 5, pp. 318–324, May 1981.

[6] J. Reed and U. Yechiali, "Queues in tandem with customer deadlines and retrials," *IEEE/ACM Trans. Netw.*, vol. 73, no. 1, pp. 1–34, Mar 2013.

[7] Y. Wang, M. C. Vuran, and S. Goddard, "Cross-layer analysis of the end-to-end delay distribution in wireless sensor networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 1, pp. 305–318, Feb 2012.

[8] D. Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris, *Fundamentals of Queueing Theory*. Wiley, 2008.

[9] A. Eckberg, "The single server queue with periodic arrival process and deterministic service times," *IEEE Trans. Commun.*, vol. 27, no. 3, pp. 556–562, March 1979.

[10] A. M. Dukhovny, "Markov chains with quasitoeplitz transition matrix," *Journal of Applied Mathematics and Simulation*, vol. 2, no. 1, pp. 71–82, 1989.

[11] W. K. Grassmann and D. P. Heyman, "Equilibrium distribution of block-structured Markov chains with repeating rows," *J. Appl. Prob.*, vol. 27, pp. 557–576, 1990.

[12] Y. Q. Zhao, W. Li, and W. J. Braun, "Censoring, factorizations, and spectral analysis for transition matrices with block-repeating entries," *Methodology and Computing in Applied Probability*, vol. 5, no. 1, pp. 35–58, 2003.

[13] U. N. Bhat, *An Introduction to Queueing Theory*. Birkhuser, 2008.

[14] J. A. Ferreira, "Pairs of renewal processes whose superposition is a renewal process," *Stochastic Processes and their Applications*, vol. 86, no. 2, pp. 217–230, 2000.