Analytical Approximation of Packet Delay Jitter in Simple Queues

Karim Hammad, Abdallah Moubayed, Abdallah Shami, and Serguei Primak

***Abstract*—Delay jitter of data packets is known to be a crucial quality of service measure especially for real-time applications (e.g., VoIP). It takes place as a result of the queuing, scheduling, and routing latencies within the network. However, control schemes that directly tackle the jitter problem in today's advanced wireless systems are rare. To enable such schemes, proper modeling of the packet delay jitter is an essential preliminary step. This letter presents a comprehensive mathematical modeling for the packet delay jitter in a simple queuing system with one traffic buffer of infinite length, one server, and single hop. In contrast to independent and identically distributed models, our analysis focuses on the correlated nature of service intervals. The presented models study different scenarios and parameters for the queue in terms of the system's utilization and the probability distribution of data packets' service and interarrival times, respectively. Numerical simulations demonstrate the high accuracy achieved by the presented models.**

***Index Terms*—Delay jitter, QoS, queuing theory.**

I. Introduction

The UNABATED evolution of today's wireless technologies, which support extremely high data rates, becomes evidently the main driver of the current wide spectrum of multimedia services. This spectrum ranges from services such as VoIP, real-time video streaming, social networking, interactive on-line gaming and ends up by the new evolving concept of Internet of Things (IoT). The emergence of such services over the currently deployed 4G or even the future 5G networks is associated with stringent QoS requirements. One vital QoS metric that substantially affects the end-to-end experience of real-time services is the packet delay jitter.

In literature, various attempts have been carried out either to control the delay jitter [1], [2] or to provide an analytical approximation to it [3], [4]. Houeto and Pierre [1] developed a jitter constraint admission control mechanism to provide eventual guarantees on the delay jitter bounds in multi-service ATM networks. Utilizing a different approach, Oklander and Sidi [2] provided an analytical model for the well known jitter buffer mechanism. The jitter buffer mechanism is known as a post-processing scheme implemented at the receiver side to compensate for the delay jitter encountered by the packet throughout its network route. In contrast to Houeto's approach, Matragi *et al.* [3] proposed a probabilistic model for estimating the end-to-end jitter of a periodic traffic (by means of estimating the departure process) traversing through multiple nodes in an ATM network. Similarly, Wen *et al.* [4] provided a theoretical evaluation for the packet delay jitter of a real-time service in double queue single server with limited capacity queuing system. The real-time traffic flow was modeled as Two-state Markov-modulated Bernoulli process (MMBP-2) while the other non real-time flow was modeled as an interrupted Bernoulli process (IBP).

Considering the simplest queuing model with a single flow per node server and single hop for each packet, the packet delay jitter is due to the stochastic nature of both of the packets evolution and the channel quality serving these packets. In this context and to the best of our knowledge, it has been noted that none of the published research has provided a clear mathematical expressions describing the jitter behavior. As a result, this letter is devoted to presenting a solid mathematical characterization for the delay jitter of the simple queuing model in different scenarios. The underlying objective is to help derive heuristic algorithms for a jitter-efficient packet scheduler that could be utilized in today's wireless networks.

The rest of the letter is organized as follows. Section II presents the jitter model for the single-flow single-server queuing system under different traffic loads, and packets' interarrival and service time statistics. The numerical validation for the derived models in Section II are provided in Section III. Finally, Section IV concludes the letter.

II. Analytical Approximation of Jitter in Queues

Following [3], we define jitter $\Delta t(k+1,k) = \tau_{k+1,d} - \tau_{k,d}$ as a time difference between delays experienced by two sequential packets indexed as k and $k+1$, where $\tau_{k,d}$ is the queuing delay for packet indexed k. At this point the values of $\Delta t(k+1,k)$ could be negative, in spite of the standard practice of considering only the absolute value of the jitter. In this section we will be interested in approximating the distribution of $\Delta t(k+1,k)$ and, above all, the approximation of its mean $m_j = \mathcal{E}\{\Delta t(k+1,k)\}$, absolute value mean $m_{|j|} = \mathcal{E}\{|\Delta t(k+1,k)|\}$ and the variance $\sigma_j^2 = \mathcal{E}\{(\Delta t(k+1,k))^2\}$. In general, the problem is hard to solve analytically. In order to simplify the analysis we consider three modes of queue operation based on the system's utilization factor ρ

1) underloaded (underutilized) system $\rho \ll 1$
2) critically loaded system $1-\rho \ll 1$
3) intermediate case.

A. Underutilized Queue

In the underutilized system the queue is very shallow, consisting mainly of a single packet. In other words, as soon as a packet gets into the queue, it starts being served. The difference in the delay between two sequential packets in underutilized queue is thus defined only by a difference in the service time (and independent of the arrival process), *i.e.,*

$$\Delta t(k+1,k) = \tau_{k+1,s} - \tau_{k,s} \tag{1}$$

where $\tau_{k,s}$ is the time needed to service k-th packet. As a result, the statistics of $\Delta t(k+1,k)$ is defined by the joint distribution

Manuscript received August 6, 2016; accepted August 15, 2016. Date of publication August 19, 2016; date of current version December 15, 2016. The associate editor coordinating the review of this paper and approving it for publication was S. Zhou.

The authors are with Western University, London, ON N6A 3K7, Canada (e-mail: khammad@uwo.ca; amoubaye@uwo.ca; ashami2@uwo.ca; slprimak@uwo.ca).

Digital Object Identifier 10.1109/LWC.2016.2601609

of two sequential service times $p_{2s}(\tau_1, \tau_2)$. Using well known results from statistics [5], the distribution of difference of two random variables is given as

$$p_j(z) = \int_{-\infty}^{\infty} p_{2s}(\tau_1, \tau_1 + z) d\tau_1 \quad (2)$$

In some cases, such as *G/M/1* queues, the service time for different packets are independent, which allows for a simpler expression of the jitter distribution:

$$p_j(z) = \int_{-\infty}^{\infty} p_s(\tau_1) p_s(\tau_1 + z) d\tau_1 \quad (3)$$

i.e., it depends only on the marginal PDF $p_s(\tau)$ of the service time. It is worth noting here that for deterministic constant service time $p_s(\tau) = \delta(\tau - T_s)$ one immediately obtains $p_j(z) = \delta(z)$, *i.e.,* in the case of a light load there is no jitter introduced by the queue (*i.e.,* arrival process) and the server combined.

1) G/M/1 Queue: In this case, the service time is exponentially distributed with the average rate μ packets per second

$$p_s(\tau) = \mu \exp(-\mu\tau) u(\tau) \quad (4)$$

Here $u(\tau)$ is the Heaviside unit step function [6]. Making use of equation (3) one obtains

$$\begin{aligned} p_j(z) &= \mu^2 \int_{-\infty}^{\infty} \exp(-\mu\tau_1) \exp(-\mu\tau_1 - \mu z) u(\tau_1) u(\tau_1 + z) d\tau_1 \\ &= \frac{\mu}{2} \exp(-\mu|z|) \end{aligned} \quad (5)$$

Using the PDF (5) one can obtain the following expressions for the mean, absolute mean and the variance of jitter

$$m_j = 0,\ m_{|j|} = \frac{1}{\mu},\ \sigma_J^2 = \frac{2}{\mu^2} \quad (6)$$

Thus, in order to reduce jitter, one has to increase the service rate μ.

2) Gilbert-Elliot (GE) Channel: Let us assume that packets of a fixed length L bits are being served by a Gilbert-Elliot channel with the service rate R_B in the "BAD" state B, and the rate $R_G > R_B > 0$ in the "GOOD" state G. The transition between the states are described by the following transition matrix [7]

$$\mathbf{T} = \begin{bmatrix} (1-d)P_G + d & (1-d)(1-P_G) \\ (1-d)P_G & (1-d)(1-P_G) + d \end{bmatrix} \quad (7)$$

In this model P_G is the probability of the "GOOD" state, while $0 \le d \le 1$ defines the correlation properties of the channel. If $d = 0$ the states are changing in an independent manner, while $d = 1$ corresponds to the situation of no transition to another state (constant channel). The service time of a packet in the state G is $\tau_G = L/R_G$, while in the state B is given by $\tau_B = L/R_B$. Considering two sequential packets, one can observe that no jitter will appear if the channel does not change its state (*i.e., GG* or *BB* combination appears), while the difference $\tau_G - \tau_B = -\tau_J$ corresponds to the channel changing from B to G and $\tau_B - \tau_G = \tau_J$ corresponds to GB transition. Here

$$\tau_J = \frac{L}{R_B} - \frac{L}{R_G} = \frac{L(R_G - R_B)}{R_G R_B} \quad (8)$$

Therefore, the distribution of jitter is given as

$$\begin{aligned} p_j(z) &= (1 - P_{GB} - P_{BG})\delta(z) + P_{GB}\delta(z - \tau_J) + P_{BG}\delta(z + \tau_J) \\ &= [1 - 2(1-d)P_G(1-P_G)]\delta(z) + (1-d)P_G(1-P_G) \\ &\quad \times \delta(z - \tau_J) + (1-d)P_G(1-P_G)\delta(z + \tau_J) \end{aligned} \quad (9)$$

Using the PDF (9) one can obtain the following expressions for the mean, absolute mean and the variance of jitter

$$\begin{aligned} &m_j = 0,\ m_{|j|} = 2(1-d)\tau_J P_G(1-P_G), \\ &\sigma_J^2 = 2\tau_J^2(1-d)P_G(1-P_G) \end{aligned} \quad (10)$$

It can be seen from (10) that the reduction of jitter could be achieved in a number of ways:

- Equalizing the service time in each state of the channel (reduction of τ_J) by the channel inversion [8].
- Increasing speed of a mobile leads to lowering of d, thus it produces an increased jitter. Converse will reduce jitter.
- Increasing the probability of one state over the other will lead to reduction of jitter. This could also be achieved by channel inversion.

3) Correlated Exponential Service Time: There is no single model for the bivariate exponential distribution, even in the Markov case [7]. However, to investigate the effect of service time correlation, we choose a simple distribution, suggested in [9] for exponentially correlated Markov processes

$$\begin{aligned} p_{2s}(\tau_1, \tau_2) &= (1-d)\mu^2 \exp[-\mu(\tau_1 + \tau_2)] u(\tau_1) u(\tau_2) \\ &\quad + d\mu \exp(-\mu\tau_1) u(\tau_1) \delta(\tau_1 - \tau_2) \end{aligned} \quad (11)$$

where τ_1 and τ_2 represent two consecutive and correlated service times, and d is the correlation parameter as in (7). Simple algebra results into the following expression for the jitter distribution:

$$p_j(z) = (1-d)\frac{\mu}{2} \exp(-\mu|z|) + d\delta(z) \quad (12)$$

Therefore, the mean and the variance are given as:

$$m_j = 0,\ m_{|j|} = (1-d)\frac{1}{\mu},\ \sigma_J^2 = (1-d)\frac{2}{\mu^2} \quad (13)$$

Not surprisingly, this example confirms the conclusion of Section II-A2 that correlation suppresses jitter.

B. Heavy Loaded Queue

In the case of a heavy load $1 - \rho \ll 1$ the queue is almost never empty. Let us assume that the k-th packet starts being served at time instant $t = 0$. Let $t_0 < 0$ be a time of arrival of the k-th packet to the queue. According our assumption of a long queue, the time $t_0 + \tau_A$ of arrival of the $k+1$-th packet, also precedes $t = 0$, *i.e.,* this packet is in the queue by the beginning of service time of the k-th packet. Here τ_A is the interarrival time. The k-th packet will be served at the time instant $t = \tau_{k,s}$, while the $k+1$-th packet will be served at the time instant $t = \tau_{k,s} + \tau_{k+1,s}$. Therefore, the jitter $\Delta t(k+1, k)$ could be evaluated as

$$\begin{aligned} \Delta t(k+1, k) &= (\tau_{k+1,s} + \tau_{k,s} - t_0 - \tau_A) - (\tau_{k,s} - t_0) \\ &= \tau_{k+1,s} - \tau_A \end{aligned} \quad (14)$$

Thus, the jitter is now a function of the arrival and service time distributions.

1) M/M/1 Queue: Let us assume that packets are arriving at a rate λ and being served at the rate $\mu > \lambda = \mu\rho$. In this case both interarrival time and service time are exponentially distributed, *i.e.,*

$$p_A(\tau) = \lambda \exp(-\lambda\tau)u(\tau),\ p_s(\tau) = \mu \exp(-\mu\tau)u(\tau) \quad (15)$$

Following equation (2) one can derive the following distribution of the jitter $\Delta t(k+1, k)$

$$p_j(z) = \frac{\mu\lambda}{\mu+\lambda}\begin{cases}\exp(-\lambda z) & \text{if } z \geq 0\\ \exp(\mu z) & \text{if } z < 0\end{cases} \quad (16)$$

The corresponding mean, absolute mean and the variance are given by

$$m_j = \frac{\mu-\lambda}{\lambda\mu} = \frac{1}{\mu}\frac{1-\rho}{\rho} \approx 0,\ \ m_{|j|} = \frac{1}{\mu}\frac{1+\rho^2}{(1+\rho)\rho} \approx \frac{1}{\mu},$$
$$\sigma_j^2 = \frac{1}{\mu^2}\frac{1+\rho^2}{\rho^2} \approx \frac{2}{\mu^2} \quad (17)$$

It can be seen that for the heavy loaded system, the impact on jitter could be similar to that of the light loaded system. It is important to note that this is true only in the case of *M/M/1* queue.

2) D/M/1 Queue: In the case of deterministic interarrival intervals $p_A(\tau) = \delta(\tau - T_A)$ where $T_A = 1/\lambda = 1/\rho\mu$ is the interarrival interval. Therefore, the distribution of jitter is just a shifted version of the service time distribution

$$p_j(z) = \mu \exp[-\mu(z + T_A)]u(z + T_A) \quad (18)$$

The load of the system $\rho = 1/T_A\mu$. The corresponding mean, absolute mean and the variance are given by

$$m_j = T_A - \frac{1}{\mu} = \frac{1-\rho}{\rho}\frac{1}{\mu} \approx 0,$$
$$m_{|j|} = \frac{2\rho\exp(-1/\rho) + (1-\rho)}{\rho}\frac{1}{\mu},\ \ \sigma_j^2 = \frac{2}{\mu^2} \quad (19)$$

3) GE Channel: Let the server be described by a set of M states with probabilities P_m and service rate $R_m > 0$ in the m-th state. The density of the service time $\tau_{s,m} = L/R_m$ is given by

$$p_s(\tau) = \sum_{m=1}^{M} P_m\delta(\tau - \tau_{s,m}) \quad (20)$$

Furthermore, assuming that $p_A(\tau)u(\tau)$ is the PDF of the interarrival time, the distribution of the jitter could be expressed as

$$p_j(z) = \sum_{m=1}^{M} P_m p_A(-z + \tau_{s,m})u(-z + \tau_{s,m}) \quad (21)$$

It is worth noting at this stage that correlation of the underlying Markov channel does not affect the jitter distribution, since a service time for a later packet includes the whole service time of the preceding packet.

C. Bridging Case

In the intermediate case when the incoming traffic does not always keep queue occupied we suggest the following *approximation* to the distribution of jitter. Let P_0 be the probability of the empty queue. In this case

$$p_j(z) = P_0 p_{j,l}(z) + (1 - P_0)p_{j,h}(z) \quad (22)$$

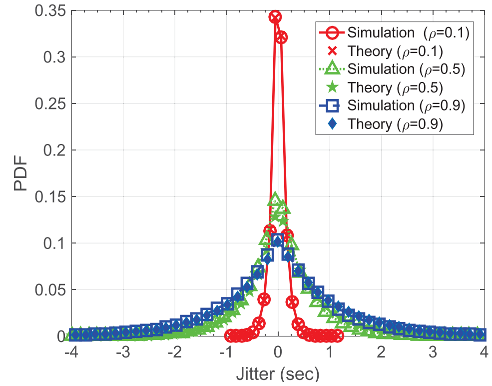

Fig. 1. Packet delay jitter evaluation for the *M/M/1* queue.

Here $p_{j,l}(z)$ is the jitter PDF in the case of low load while $p_{j,h}(z)$ is the PDF of the jitter in the case of high load. The exact value of P_0 is known in some cases of classical queues. For example, for *M/M/1* queue $P_0 = 1-\rho = 1-\lambda/\mu$, therefore one obtains

$$p_j(z) = (1-\rho)\frac{\mu}{2}\exp(-\mu|z|) + \frac{\rho^2}{1+\rho}\begin{cases}\exp(-\lambda z) & \text{if } z > 0\\ \exp(\mu z) & \text{if } z < 0\end{cases} \quad (23)$$

Other approximation could be obtained in the very much same manner. In many sources it is suggested that P_0 is substituted by $1 - \rho$ in the case of an arbitrary sources.

III. Numerical Simulations

This section presents MATLAB numerical evaluation for the packet delay jitter models developed in the previous section. Two types of traffic are considered. First, the Poisson traffic is considered in the case of *M/M/1*, *M/GE/1* and *M/ExpCorr/1* (*i.e.,* exponentially correlated service time case) queues. The second type is the deterministic traffic, considered in the *D/M/1* queue, where packet arrivals are assumed to be periodic. In both cases the packet arrival rate (*i.e.,* λ) is kept constant at 10^3 packets/sec while the average service time (*i.e.,* $1/\mu$) in msec takes the values {0.1, 0.5, 0.9} to study the system at the light, moderately and heavy loaded queue modes (*i.e.,* $\rho = \{0.1, 0.5, 0.9\}$ value), respectively.

For the *M/M/1* queue, the results depicted in Fig. 1 show a very good agreement for the derived jitter models, described in (5), (16) and (23), with the simulation results at the three queue modes. In addition, it is shown that increasing the queue load ρ (*i.e.,* packets' queuing delay) increases the packet delay jitter variance. In contrast to *M/M/1*, when setting the packets' interarrival time to a constant value (*i.e.,* $T_A = 1$ msec), the *D/M/1* queue delay jitter was found as shown in Fig. 2. The results illustrate that the analytical jitter models developed in (5), (18) and (23) provide an accurate approximation for the simulation results.

To investigate the effect of packets' service time correlation on the jitter performance of the *M/M/1* queue with light load (*i.e.,* $\rho = 0.1$), the Markov model for exponentially correlated process developed in [9] is utilized to set exponentially correlated service times for the generated packets. The results presented in Fig. 3 show the performance of the jitter model developed in (12) in comparison with the simulation results and the effect of changing the correlation parameter (*i.e.,* d)

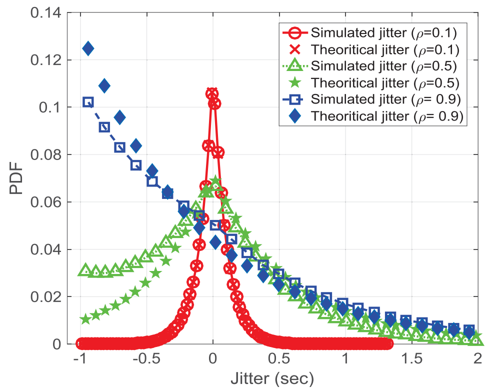

Fig. 2. Packet delay jitter evaluation for the *D/M/1* queue.

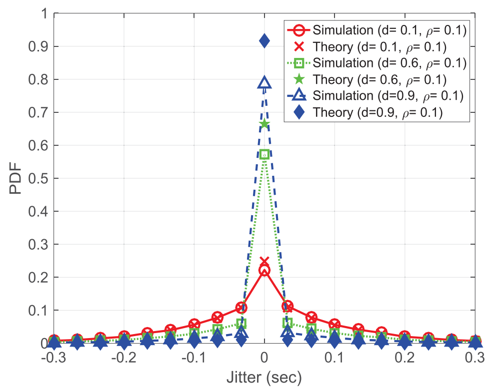

Fig. 3. Packet delay jitter evaluation for the *M/ExpCorr/1* queue.

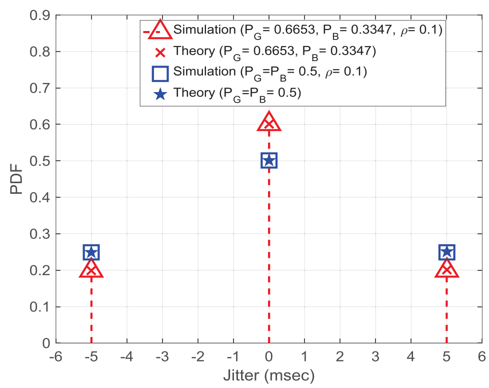

Fig. 4. Packet delay jitter evaluation for underutilized *M/GE/1* queue.

on both. It is noted that the analytical model shows a good representation for the simulation results especially at low correlation. On the other hand, increasing the packets' service time correlation sharpens the delay jitter distribution around zero (*i.e.,* the delta term in equation (12)). This behaviour pertains to the fact that high correlation for service time (*i.e.,* large value for d), while knowing the service time to be the sole factor affecting the jitter in case of low loaded system, implies high correlation for the packet delay, and hence, low packet delay jitter.

For the case of GE server, and *M/GE/1* queue, the analytical jitter models derived in (9) and (21) are numerically evaluated in Fig. 4 and Fig. 5 for low and high load, respectively.

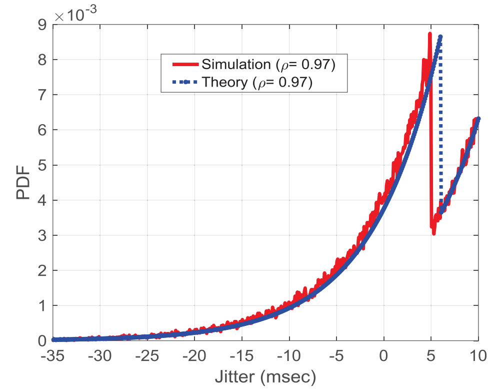

Fig. 5. Packet delay jitter evaluation for heavy loaded *M/GE/1* queue.

In the case of underutilized queue, Fig. 4 results show the discrete distribution for the jitter, as in (9), at the values 0 and $\pm\tau_j$. It should be noted that the GE model was set to have $\tau_G = 5$ msec and $\tau_B = 10$ msec for both cases. Moreover, the results shown in Fig. 4 demonstrate a subtle match between the theoretical model and simulation at different probabilities for the good and bad states. Finally, at high queue load, the jitter model suggested in (21) for the *M/GE/1* queue highly represents a real simulated queue as illustrated in Fig. 5.

IV. Conclusion

An extensive analytical model for the delay jitter in the single queue single server queuing system has been developed. It was found that at different utilization levels and service statistics for the queue, the parameters affecting the delay jitter change accordingly. The presented expressions were tested via numerical simulations and were found to be accurate. The insights gained in this letter are currently being utilized to design various jitter-efficient heuristic packet scheduling algorithms for LTE networks in a sequel to this letter.

References

[1] F. Houeto and S. Pierre, "Characterization of jitter and admission control in multiservice networks," *IEEE Commun. Lett.*, vol. 8, no. 2, pp. 125–127, Feb. 2004.

[2] B. Oklander and M. Sidi, "Jitter buffer analysis," in *Proc. 17th Int. Conf. Comput. Commun. Netw. (ICCCN)*, Aug. 2008, pp. 1–6.

[3] W. Matragi, K. Sohraby, and C. Bisdikian, "Jitter calculus in ATM networks: Multiple nodes," *IEEE/ACM Trans. Netw.*, vol. 5, no. 1, pp. 122–133, Feb. 1997.

[4] D. Wen, C. Xue-Fen, L. Zi-Chuan, and Z. Yu-Hong, "Queuing theory based analysis for packet jitter of mixed services," *J. China Univ. Posts Telecommun.*, vol. 21, no. 3, pp. 71–76, 2014.

[5] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. Boston, MA, USA: McGraw-Hill, 1991.

[6] M. Abramowitz and I. Stegun, Eds., *Handbook of Mathematical Functions*. New York, NY, USA: Dover, 1965.

[7] S. Primak, V. Kontorovich, and V. Lyandres, *Stochastic Methods and Their Applications to Communications: SDE Approach*. Chichester, U.K.: Wiley, 2004.

[8] M.-S. Alouini and A. J. Goldsmith, "Capacity of Rayleigh fading channels under different adaptive transmission and diversity-combining techniques," *IEEE Trans. Veh. Technol.*, vol. 48, no. 4, pp. 1165–1181, Jul. 1999.

[9] S. Primak, V. Lyandres, and V. Kontorovich, "Markov models of non-Gaussian exponentially correlated processes and their applications," *Phys. Rev. E*, vol. 63, no. 6, May 2001, Art. no. 061103.